Universidade de Brasília
IE - Instituto de Exatas
Departamento de Estatística

Regressão Logística Multinível:
Uma aplicação de Modelos Lineares Generalizados Mistos

Alex Luiz Martins Matheus da Rocha

Relatório Final do Projeto Final

Orientadora: Profª Maria Teresa Leão Costa

Brasília
Dezembro de 2014

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Em muitos estudos educacionais a população de interesse tem estrutura multinível, ou hierárquica, como no caso em que o interesse do estudo é avaliar determinada varíavel resposta de alunos que estão agrupados em turmas. Para essa situação, modelos hierárquicos são os mais adequados.

Modelos hierárquicos são modelos estatísticos usados para analisar dados hierárquicos, pois levam em conta as várias dependências e permitem analisar todos os níveis da hierarquia. Esse tipo de modelo é chamado de um modelo misto, pois possui tanto efeitos fixos como aleatórios.

Este trabalho apresenta uma aplicação desse tipo de modelo no estudo de fatores que influenciam o rendimento dos alunos que cursam as disciplinas de serviço do departamento de Estatística da UnB, identificando os efeitos desses fatores em cada uma das disciplinas.

Palavras-chave: Regressão Logística, Regressão Multinível, Modelos Mistos, Modelos Lineares Generalizados Mistos, Modelos Hierárquicos, Avaliação Educacional.

Abstract

In many educational studies the population of interest has a multilevel or hierarchical structure, such as when the interest of the study is to evaluate a certain response variable of students clustered within classes. Hierarchical models are better suited for this situation.

Hierarchical Models are statistical models used to analyze hierarchical data, taking into account the many dependencies and allowing the analysis of all hierarchical levels. This type of model has both fixed and random effects, making it a particular case of a mixed model.

This paper presents an application of such model to study which factors affect the performance of Statistics students at UnB, identifying their effects on different courses.

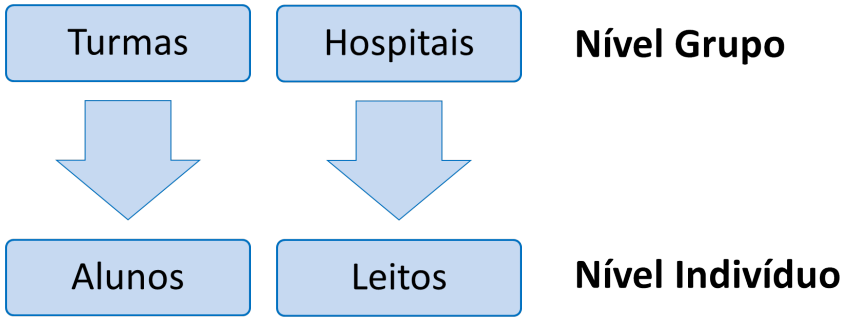
Keywords: Logistic Regression, Multilevel Regression, Mixed Models, Generalized Linear Mixed Models, Hierarchical Models, Educational Evaluation.

Capítulo 1

Introdução e Justificativa

A análise estatística por meio de modelos de regressão que explicam a variabilidade de uma variável de interesse, em função de outras, é de muita importância em diversas áreas científicas, como por exemplo na educação, onde é comum se ter o interesse de estudar sobre quais fatores influenciam no desempenho dos alunos em determinada prova (Laros et al 2010). Nesse tipo de estudo é comum que a população tenha estrutura hierárquica, como a ilustra a figura abaixo.

Exemplo de populações com estrutura hierárquica, ou multinível, de 2 níveis.



Para se aplicar o modelo de regressão usual, vários pressupostos teóricos são feitos, como independência entre as observações. Mas no caso em que a população tem estrutura hierárquica, como apresentado na figura acima, alunos dentro de uma mesma turma tendem a ter características semelhantes

devido a conviverem no mesmo ambiente, com o mesmo professor, de forma que não há independência entre esses indivíduos. Se após o ajuste do modelo esses pressupostos não forem satisfeitos, conclusões incorretas podem ser tiradas da análise dos dados, devido ao viés dos estimadores ou erro padrão das estimativas subestimado ou superestimado.

Por esse motivo, modelos que incorporam a estrutura hierárquica dos dados e as dependências entre indivíduos têm sido cada vez mais utilizados. Esses modelos são comumente chamados em outras áreas por modelos multinível ou hierárquicos, mas na Estatística, é um caso particular de uma classe de modelos chamada de modelos com efeitos mistos, de forma que toda inferência estatística que pode ser feita vem da análise dessa classe.

A importância desse modelo vem da necessidade de satisfazer os pressupostos de um modelo estatístico. Toda a análise inferencial dos dados depende dessas suposições e da forma como os dados foram obtidos. Com computadores cada vez mais rápidos e softwares mais evoluídos, esse tipo de modelo complicado pode ser facilmente ajustado, de modo a representar melhor a realidade, com inferências mais precisas.

Tendo em vista o que foi mencionado, o objetivo geral desse trabalho é estudar e aplicar a regressão multinível, bem como identificar como esse tipo de modelo se encaixa na estrutura dos modelos com efeitos mistos.

Os objetivos específicos consistem em:

- Identificar os efeitos desses fatores em cada disciplina.
- Fazer uma relação entre modelos hierárquicos e modelos com efeitos mistos, apresentando no referencial teórico desse trabalho como a regressão multinível é um caso particular de um modelo misto.
- Apresentar como o ajuste de modelos hierárquicos generalizados é feito usando o SAS, a partir de procedimentos para modelos lineares generalizados mistos.

Capítulo 2

Referencial Teórico

2.1 Modelos Lineares Generalizados

Uma das formas de analisar o padrão de associação e interação entre uma variável de interesse, denominada variável resposta, e outras variáveis, denominadas variáveis explicativas, é por meio de modelos estatísticos. Os parâmetros determinam a intensidade e a importância dos efeitos e inferências podem ser feitas sobre esses parâmetros para avaliar quais variáveis realmente afetam a resposta. Os valores preditos pelo modelo melhoram a estimativa da média da variável resposta nos possíveis valores das variáveis explicativas.

Uma classe de modelos muito utilizada em diversas análises é a de Modelos Lineares Generalizados (MLG), que é caracterizada por 3 componentes:

i) Componente Aleatório

É a especificação da distribuição de probabilidade da variável resposta Y. As observações $(y_1, ..., y_n)$ geralmente são consideradas independentes.

ii) Componente Sistemático

Especifica as variáveis explicativas, de forma linear. Sejam $(x_1, ..., x_k)$ as k variáveis explicativas. A combinação linear das variáveis explicativas,

denominada preditor linear, é dado por:

$$\alpha + \beta_1 x_1 + ... + \beta_k x_k$$

Interações entre variáveis explicativas, por exemplo $x_k = x_1 x_2$, são permitidas.

iii) Função de Ligação

Especifica a função $g(.)$ que relaciona $\mu = E(Y)$ com o preditor linear, isto é, conecta os componentes aleatório e sistemático:

$$g(\mu) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

O modelo de regressão linear e o de regressão logística, que será mencionada na próxima seção, fazem parte da classe dos MLG, considerando, respectivamente, $g(\mu) = \mu$ e Y com distribuição normal, $g(\mu) = log[\mu/(1-\mu)]$ e Y com distribuição binomial.

2.1.1 Inferência Estatística em MLG

Geralmente, as estimativas são feitas pelo método de máxima verossimilhança, utilizando as propriedades assintóticas de seus estimadores. Assim, para amostras suficientemente grandes, os intervalos de Wald de (1-α)% de confiança para os parâmetros β_j são dados por:

$$\hat{\beta}_j \pm z_{\alpha/2} SE(\hat{\beta}_j) \tag{2.1}$$

Onde $SE(\beta_j)$ é o erro padrão associado ao estimador de β_j e $z_{\alpha/2}$ é o valor tal que P(Z $\geq z_{\alpha/2}$) = α/2 , onde Z $\sim$ Normal(0,1).

Pode-se usar o teste de Wald para testar a significância de β_j:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Sob H_0, tem-se que:

$$z = \frac{\hat{\beta}_j}{SE(\beta_j)} \sim N(0,1) \tag{2.2}$$

Outro teste mais poderoso para testar as mesmas hipóteses, mesmo com amostras menores, é o teste da razão de verossimilhança, que compara o máximo da log-verossimilhança L_0 quando $\beta_j = 0$ (Modelo Reduzido, ou mais simples), com o máximo da log-verossimilhança L_1 sem impor restrições à β_j (Modelo completo, ou Saturado). A estatística do teste e sua distribuição sob hipótese nula é dada por:

$$-2(L_0 - L_1) \sim \chi^2_1 \tag{2.3}$$

Como o modelo saturado tem mais parâmetros que o modelo reduzido, $L_1 \geq L_0$.

A estatística do teste da razão de verossimilhança 2.3 de um MLG também é chamada de *Deviance*. Considerando que o modelo Saturado seja o modelo mais complexo possível, isto é, com todos os parâmetros, e que o modelo Reduzido seja o de interesse, pode-se usar o deviance para analisar a qualidade do ajuste do modelo. Entretanto, nem sempre o *Deviance* tem distribuição qui-quadrado, mas para a regressão logística isso acontece, com graus de liberdade igual a diferença do número de parâmetros em cada modelo.

Assim, o *Deviance* pode ser usado para a comparação de modelos aninhados. Considere que o modelo M_0 é um caso particular do modelo M_1, com os máximos da log-verossimilhança iguais a L_0 e L_1, respectivamente. Denotando por L_S o máximo da log-verossimilhança do modelo saturado, o teste da razão de verossimilhança para testar se M_1 não é significativamente melhor que M_0 é dado por:

H_0 : O modelo mais simples (M_0) se ajusta tão bem quanto (M_1).

H_1 : O modelo (M_1) se ajusta significativamente melhor que (M_0).

Sob H_0, a estatística do teste é:

$$\begin{aligned} -2[L_0 - L_1] &= -2[L_0 - L_1 + L_S - L_S] \\ &= -2[L_0 - L_S - L_1 + L_S] \\ &= -2[L_0 - L_S] - (-2)[L_1 - L_S] \\ &= D_0 - D_1 \sim \chi^2_g \end{aligned} \tag{2.4}$$

Onde D_0 e D_1 são, respectivamente, os *deviances* referentes ao modelo M_0 e M_1 e g $>$ 0 é a diferença do número de parâmetros nesses modelos. Quanto maior a diferença entre D_0 e D_1, maior é a evidência de que modelo M_1 se ajusta melhor do que M_0.

Para comparar modelos não aninhados, isto é, um não é um caso particular do outro, usamos as medidas AIC ou BIC, que são respectivamente o critério de informação de Akaike (Akaike, 1974), e o critério de informação Bayesiano (Schwarz, 1978). Seja L o máximo da verossimilhança de um determinado modelo, p seu número de parâmetros e n o tamanho da amostra, os critérios de informação são dados por:

$$AIC = -2ln(L) + 2p \tag{2.5}$$

$$BIC = -2ln(L) + pln(n) \tag{2.6}$$

Quanto menor os valores de 2.5 e 2.6, melhor o ajuste do modelo. Entretanto como não temos distribuição de probabilidade envolvida, não sabemos quando um valor AIC ou BIC é significativamente melhor do que outro. Se

os modelos de comparação forem aninhados, no sentido de um deles ser um caso particular do outro, é melhor usar o teste da razão de verossimilhança, apresentado anteriormente.

2.2 Regressão Logística

Regressão Logística é um modelo estatístico usado quando se deseja explicar uma variável resposta categórica em função de variáveis explicativas quantitativas ou qualitativas. Será considerado apenas o caso em que a resposta é binária, ou seja, a variável possui dois possíveis resultados, ou categorias : sucesso ou fracasso. Nesse modelo, a probabilidade de sucesso depende de outras variáveis. O termo regressão logística simples refere-se ao caso em que tem-se apenas uma variável explicativa.

Um conceito muito importante nesse tipo de modelagem é o de chance, usualmente chamado pelo nome em ingles, *odds*. Suponha que temos uma variável Y com duas categorias: sucesso, denotado por 1, ou fracasso, denotado por 0. Para a probabilidade de sucesso P(Y $=$ 1) $= \pi$, a chance de sucesso é definida por:

$$odds = \frac{\pi}{(1-\pi)} \tag{2.7}$$

Se a probabilidade de sucesso é maior que a de fracasso, *odds* > 1. Caso contrário, *odds* < 1. Se, por exemplo, *odds* $=$ 2 , a probabilidade de sucesso é duas vezes a probabilidade de fracasso. Por outro lado, se *odds* $=$ 0.5, a probabilidade de sucesso é metade da probabilidade de fracasso. Isolando π na equação 2.7 obtemos que:

$$\pi = \frac{odds}{odds+1} \tag{2.8}$$

Em tabelas de contingência 2x2 com duas variáveis binárias X e Y, a razão

de duas chances, chamada de razão de chances, ou no inglês *odds ratio*, é dada por:

$$\theta = \frac{odds_1}{odds_2} = \frac{\frac{\pi_1}{(1-\pi_1)}}{\frac{\pi_2}{(1-\pi_2)}} \tag{2.9}$$

Onde $odds_1$ é a chance de sucesso para a variável Y na categoria 1 de X e $odds_2$ é a chance de sucesso para Y na categoria 2 de X. Se $\theta > 1$, a chance de sucesso na categoria 1 de X é maior que a chance de sucesso na categoria 2 de X. Analogamente, se $\theta < 1$, a chance de sucesso na categoria 1 de X é menor do que na outra categoria.

2.2.1 Regressão Logística Simples

Suponha que a variável resposta Y tenha duas categorias como descrito anteriormente e que a única variável explicativa X seja quantitativa. Seja $\pi(x)$ a probabilidade de sucesso de Y no valor x, o modelo de regressão logística tem forma linear para o logito dessa probabilidade:

$$logito(\pi(x)) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \tag{2.10}$$

Ou seja, o logarítmo natural da chance de sucesso de Y no valor x tem forma linear. Dessa forma, a probabilidade de sucesso é obtida isolando $\pi(x)$ em 2.10:

$$exp(\alpha + \beta x) = \frac{\pi(x)}{(1-\pi(x))}$$

$$\frac{(1-\pi(x))}{\pi(x)} = \frac{1}{exp(\alpha + \beta x)}$$

$$\frac{1}{\pi(x)} = \frac{1 + exp(\alpha + \beta x)}{exp(\alpha + \beta x)}$$

Logo obtemos:

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)} = \frac{1}{1 + exp(-\alpha - \beta x)} \tag{2.11}$$

Em 2.10, o parâmetro β indica que para um acréscimo de 1 unidade em x, o logito aumenta em β unidades. Já em 2.11, tem-se que β determina a taxa de crescimento ou decrescimento da curva para $\pi(x)$, que tem um formato de S. Uma medida importante nessa curva é o nível mediano efetivo, denotado por $EL_{(50)}$, que indica a que nível de x a probabilidade de cada resposta de Y é 50%. Essa medida é obtida fazendo:

$$\frac{1}{2} = \frac{1}{1 + exp(-\alpha - \beta x)}$$

$$exp(-\alpha - \beta x) = 1$$

Aplicando logarítmo e isolando x,

$$EL_{(50)} = -\frac{\alpha}{\beta} \tag{2.12}$$

Quando $\hat{\beta} > 0$, a probabilidade estimada $\hat{\pi}$ é maior para maiores valores de x. Analogamente, se $\hat{\beta} < 0$, $\hat{\pi}$ é menor para maiores valores de x.

O *odds* e o *odds ratio* podem ser rapidamente obtidos no modelo de regressão logística:

$$\frac{\pi(x)}{(1 - \pi(x))} = exp(\alpha + \beta x) = e^{\alpha}\left(e^{\beta}\right)^{x} \tag{2.13}$$

O *odds ratio*, ou razão de chances, entre o nível x + 1 e o nível x é dado por:

$$\frac{\dfrac{\pi(x+1)}{(1 - \pi(x+1))}}{\dfrac{\pi(x)}{(1 - \pi(x))}} = \frac{e^{\alpha}\left(e^{\beta}\right)^{(x+1)}}{e^{\alpha}(e^{\beta})^{x}} = exp(\beta x + \beta - \beta x)$$

$$\theta = \frac{\frac{\pi(x)}{(1-\pi(x))}}{\frac{\pi(x+1)}{(1-\pi(x+1))}} = e^{\beta} \tag{2.14}$$

Em 2.13, tem-se que a chance de sucesso é multiplicada por e^{β} para cada acréscimo de 1 unidade em x. Já em 2.8 tem-se que a razão das chances de sucesso de Y em x + 1 e de Y em x, ou o *odds ratio*, é dado por e^{β}. Ou seja, a chance de sucesso no nível x + 1 é e^{β} vezes a chance de sucesso no nível x.

Suponha agora que X é uma variável categórica com 2 categorias, sucesso ou fracasso, com valores respectivamente 1 e 0. Assim, o modelo ainda é o mesmo que em 2.10, e a interpretação do *odds ratio* é parecida : a chance de sucesso de Y na categoria X $= 1$ é e^{β} vezes a chance de sucesso na categoria X $= 0$.

Se X for categórica com $C > 2$ categorias, existe a necessidade de introduzir C-1 variáveis comumente chamadas de dummy, que são variáveis indicadoras da categoria.

2.2.2 Regressão Logística Múltipla

No caso em que se tem mais de uma variável explicativa, a regessão é usualmente chamada de regressão múltipla. Suponha que X é uma variável categórica com C categorias. Usando a seguinte codificação para as C-1 variáveis dummy, com $j = 1, ..., n$:

$$X_j = \begin{cases} 1 & \text{se pertence à categoria j;} \\ 0 & \text{Caso contrário.} \end{cases}$$

O modelo é dado por:

$$logito(\pi(x_1, ..., x_{C-1})) = \alpha + \sum_{j=1}^{C-1} \beta_j x_j \tag{2.15}$$

Assim, para a categoria j, $x_i = 0$ para i $\neq$ j, e o modelo se torna:

$$logito(\pi(x_1 = 0, ..., x_j = 1, ..., x_{C-1} = 0)) = \alpha + \beta_j x_j \tag{2.16}$$

A razão de chances entre a categoria j e a categoria $k \neq j$ é obtida da mesma forma como foi feito em 2.14:

$$\theta = \frac{exp(\alpha + \beta_j)}{exp(\alpha + \beta_k)} = \frac{exp(\beta_j)}{exp(\beta_k)}$$

$$\theta = exp(\beta_j - \beta_k) \tag{2.17}$$

Dessa forma, a chance de sucesso na categoria j é e^{β_j} vezes a chance de sucesso na categoria C, que é a categoria de referência obtida quando $x_1, ..., x_{C-1} = 0$. Comparando outras categorias, a chance de sucesso na categoria j é $exp(\beta_j - \beta_k)$ vezes a chance de sucesso na categoria k.

No caso geral, que se tem como resposta tanto variáveis quantitativas quanto qualitativas com qualquer número de categorias, a notação é a mesma de 2.15.

Os modelos descritos até aqui não incluem interação. Suponha agora que X_1 é uma variável quantitativa e X_2 é uma variável categórica com 2 categorias. O modelo sem interação é:

$$logito(\pi(x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{2.18}$$

Nesse modelo, a chance de sucesso na categoria 1 de X_2 é e^{β_2} vezes a chance de sucesso na outra categoria, para qualquer valor de x. No entanto, incluindo a interação obtemos:

$$logito(\pi(x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \tag{2.19}$$

Quando $x_2 = 1$, 2.13 se torna:

$$logito(\pi(x_1, x_2 = 1)) = (\alpha + \beta_2) + (\beta_1 + \beta_{12})x_1 = \alpha^* + \beta^* x_1 \tag{2.20}$$

E quando $x_2 = 0$,

$$logito(\pi(x_1, x_2 = 0)) = \alpha + \beta_1 x_1 \tag{2.21}$$

Em 2.20 e 2.21 tem-se que tanto o intercepto quanto a inclinação são diferentes, para diferentes valores de x_2. Dessa maneira, a chance de sucesso de x_1 depende da categoria x_2 e consequentemente a razão de chances entre $x_1 + 1$ e x_1 também dependem.

2.2.3 Estimação

Foi mencionado na seção anterior que o modelo de regressão logística é um MLG. Sendo assim, os testes e intervalos de confiança apresentados podem ser usados. Dessa forma, o intervalo de (1-α)% de confiança para a razão de chances θ é dado por:

$$exp(\hat{\beta}_j \pm z_{\alpha/2} SE(\beta_j)) \tag{2.22}$$

Vale observar que se o intervalo para β_j contém o número 0, tem-se que o parâmetro não é significativamente diferente de 0. Nesse caso, devido à 2.22, a razão de chances dada por e^{β_j} não é estatisticamente diferente de 1, o que indica que a chance de sucesso não depende de x_j

Para estimar as probabilidades, de 2.11, tem-se que:

$$\hat{\pi}(x) = \frac{exp(\hat{\alpha} + \hat{\beta}x)}{1 + exp(\hat{\alpha} + \hat{\beta}x)} \tag{2.23}$$

Os intervalos de confiança para a estimativa da probabilidade de sucesso

usam a matriz de variância e covariância das estimativas dos parâmetros do modelo, pois primeiro se faz o intervalo de (1-α)% de confiança para o logito de $\pi(x)$, que para amostras grandes é dado por:

$$\text{Var}(\hat{logito}(\pi(\text{x}))) = \text{Var}(\hat{\alpha} + \hat{\beta}\text{x}) = \text{Var}(\hat{\alpha}) + x^2\text{Var}(\hat{\beta}) + 2\text{Cov}(\hat{\alpha},\hat{\beta}\text{x})$$

$$(\hat{\alpha} + \hat{\beta}x) \pm z_{\alpha/2}\sqrt{Var(\hat{\alpha} + \hat{\beta}x)} \tag{2.24}$$

Substituindo os limites superior e inferior de 2.24 em 2.23, obtém-se os limites superior e inferior da probabilidade.

2.3 Regressão Multinível

O termo multinível refere-se à dados estruturados de forma hierárquica, como por exemplo alunos dentro de turmas ou leitos dentro de hospitais, como foi apresentado na introdução desse trabalho. Sendo assim, a análise multinível consiste em examinar relações entre variáveis medidas em diferentes níveis de dados com essa estrutura.

Para utilizar o modelo de regressão multinível, é necessário ter, além de uma estrutura hierárquica da população, uma variável resposta medida no menor nível. Pode-se ter variáveis explicativas em quaisquer um dos níveis. Se a variável resposta for quantitativa, o modelo adequado é o de regressão linear multinível ou apenas regressão multinível,mas se a variável resposta for qualitativa, o modelo adequado é o de regressão logística multinível.

O nome regressão multinível geralmente é usado em outras áreas, como por exemplo educação e saúde. Já em Estatística esse tipo de modelo é denominado Modelo com Efeitos Mistos, ou simplesmente Modelo Misto. A parte inferencial será apresentada na próxima seção para ambos os modelos linear multinível e logística multinível.

2.3.1 Regressão Linear Multinível

Considerando o caso de uma estrutura hierárquica com 2 níveis, assumindo que temos P variáveis explicativas x no menor nível, indicadas por p, onde tem-se que p $=$ 1,...,P. Similarmente, temos Q variáveis explicativas w no maior nível, indicadas por q, com q $=$ 1,...,Q. A variável resposta, que está no menor nível, é denotada por y. Além disso, i=1,...,I, é o número de grupos ou *clusters*, e j = 1,...,n_i é o número de observações em cada grupo. A equação do modelo completo é dada por:

$$y_{ij} = \mu + \sum_{p=1}^{P} \beta_p x_{pij} + \sum_{q=1}^{Q} \gamma_q w_{qi} + \sum_{p=1}^{P} \sum_{q=1}^{Q} \theta_{pq} x_{pij} w_{qj} + \sum_{p=1}^{P} \tau_{pi} x_{pij} + G_i + e_{ij} \quad (2.25)$$

Claramente o modelo completo possui muitos parâmetros. Em regressão multinível é ainda mais importante que sejam incluídas apenas as variáveis e interações que forem de fato importantes para o estudo, pois esse tipo de modelo pode ficar facilmente super parametrizado. Em 2.25, temos:

i) Efeitos fixos:

   μ é o intercepto.

   β_p é o coeficiente de regressão das variáveis explicativas no menor nível.

   γ_q é o coeficiente de regressão das variáveis explicativas no maior nível.

   θ_{pq} é o coeficiente de regressão da interação entre níveis (*Cross-level interaction*).

ii) Efeitos aleatórios:

   τ_{pi} é o termo aleatório que indica se o coeficiente de regressão para o preditor x_p varia entre grupos.

   G_i é o erro no maior nível, que indica diferença entre grupos.

e_{ij} é o erro no menor nível, que indica diferença dentro de grupos.

iii) Suposições do modelo:

Suposição inicial é de relação linear entre a variável resposta e as explicativas.

$e_{ij} \sim N(0, \sigma_E^2)$ com variância constante (Homocedasticidade).

τ_{pi} e G_i são independentes de e_{ij} e tem distribuição normal multivariada com média 0. Var(G_i) = σ_G^2 é a variância do erro entre grupos. Var(τ_{pi}) = $\sigma_{\tau_p}^2$ é a variância dos coeficientes de regressão entre grupos. Em geral, as covariâncias entre G_i e τ_{pi} não são assumidas 0.

Modelos multinível são necessários pois quando os dados tem estrutura hierárquica, indivíduos dentro de um grupo tendem a ter características semelhantes e assim a amostra passa a não ser independente. Essa dependência pode ser expressa pelo coeficiente de correlação intra-classe ρ. Uma das formas de estimá-lo é considerar o modelo sem variáveis explicativas, que é um caso particular de um modelo de Componentes da Variância:

$$y_{ij} = \mu + G_i + e_{ij} \tag{2.26}$$

Tem-se que 2.26 não explica a variabilidade de y, apenas a decompõe em 2 termos independentes: e_{ij} com variância σ_E^2 e G_i com variância σ_G^2, que são chamados de componentes da variância.

Dessa forma, o coeficiente de correlação intra-classe ρ é dado por:

$$\rho = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} \tag{2.27}$$

Assim, ρ é a proporção da variância explicada pela estrutura de agrupamento na população. Também pode ser interpretado como a correlação esperada entre dois indivíduos escolhidos aleatoriamente dentro do mesmo grupo.

As interpretações dos parâmetros do modelo 2.25 são análogas à regressão linear múltipla, onde, por exemplo, para cada unidade da variável x_p temos um aumento (no caso do coeficiente β_p ser positivo e não termos interação) de y em média de β_p unidades, mantidas as demais variáveis constantes. Na presença de interação, a interpretação depende de outra variável explicativa.

Em regressão linear, o coeficiente de determinação, ou o coeficiente de determinação parcial, R^2, indica a proporção da variância explicada pelas variáveis explicativas. Em regressão linear multinível, esse coeficiente não tem interpretação simples, pois tanto o menor quanto o maior nível explicam parte da variação, mas existem variações do R^2 para o caso multinível (veja Hox, 2010).

A extensão para 3 ou mais níveis é análoga, mas bastante complicada, especialmente pelo fato de termos muitos parâmetros. Nesse caso, coeficientes de regressão das variáveis explicativas no menor nível podem variar em ambos segundo e terceiro níveis, além de que os coeficientes de regressão do segundo nível podem variar no terceiro nível.

Para evitar um modelo super parametrizado, não se deve incluir interações de alta ordem, à não ser que isso seja importante no estudo.

Pode-se também definir o coeficiente de correlação intra-classe. Considerando o modelo com 3 níveis e sem variáveis explicativas, sendo que agora i indica o número de grupos no terceiro nível, j indica o número de grupos no segundo nível e k o número de observações.

$$y_{ijk} = \mu + G_{1i} + G_{2j} + e_{ijk} \tag{2.28}$$

Onde agora G_{1i} representa o erro do terceiro nível com variância $\sigma^2_{G_1}$. Assim, temos 2 métodos para calcular ρ, ambos corretos (Algina, 2000):

i) Método 1 (Davis $\&$ Scott, 1995)

Para o nível 2,:

$$\rho_2 = \frac{\sigma^2_{G_2}}{\sigma^2_{G_1} + \sigma^2_{G_2} + \sigma^2_E} \tag{2.29}$$

E para o nível 3:

$$\rho_3 = \frac{\sigma^2_{G_1}}{\sigma^2_{G_1} + \sigma^2_{G_2} + \sigma^2_E} \tag{2.30}$$

ii) Método 2(Siddiqui, Hedeker, Flay $\&$ Hu, 1996)

Para o nível 2:

$$\rho_2 = \frac{\sigma^2_{G_1} + \sigma^2_{G_2}}{\sigma^2_{G_1} + \sigma^2_{G_2} + \sigma^2_E} \tag{2.31}$$

E para o nível 3:

$$\rho_3 = \frac{\sigma^2_{G_1}}{\sigma^2_{G_1} + \sigma^2_{G_2} + \sigma^2_E} \tag{2.32}$$

O método 1 identifica a proporção da variância explicada no nível 2 e nível 3. O segundo método representa uma estimativa para a correlação esperada entre dois elementos escolhidos aleatoriamente dentro do mesmo grupo. O método à ser utilizado depende da interpretação desejada.

2.3.2 Regressão Logística Multinível

Como foi mencionado anteriormente, quando a variável resposta é categórica e os dados têm estrutura hierárquica, o modelo adequado é o de regressão logística multinível. Este modelo é muito parecido com o modelo de regressão logística, incluindo os efeitos aleatórios e variáveis explicativas dos demais níveis. O modelo para o caso com 2 níveis e variável resposta binária é:

$$logito(\pi_{ij}) = \mu + \sum_{p=1}^{P} \beta_p x_{pij} + \sum_{q=1}^{Q} \gamma_q w_{qi} + \sum_{p=1}^{P} \sum_{q=1}^{Q} \theta_{pq} x_{pij} w_{qj} + \sum_{p=1}^{P} \tau_{pi} x_{pij} + G_i \tag{2.33}$$

Onde π_{ij} é a probabilidade de sucesso do indivíduo j no grupo i. Como

é usual na notação de regressão logística, 2.33 não apresenta o termo e_{ij}. As interpretações do modelo são análogas aquelas discutidas em regressão logística.

Esse modelo é um caso de modelo linear generalizado misto, que tem estrutura parecida com MLG:

i) Componente Aleatório

y_{ij} com distribuição binomial(n_{ij}, $E(y_{ij}) = \pi_{ij}$). Onde $n_{ij} = 1$, ou seja, $y_{ij} \sim$ Bernoulli(π_{ij}).

ii) Componente Sistemático

$$\mu + \sum_{p=1}^{P} \beta_p x_{pij} + \sum_{q=1}^{Q} \gamma_q w_{qi} + \sum_{p=1}^{P} \sum_{q=1}^{Q} \theta_{pq} x_{pij} w_{qj} + \sum_{p=1}^{P} \tau_{pi} x_{pij} + G_i$$

iii) Função de Ligação

$$g(\pi_{ij}) = logito(\pi_{ij}) = \frac{\pi_{ij}}{1 - \pi_{ij}}$$

Cabe observar que a variância é função da proporção populacional π_{ij}, isto é, $\sigma_E^2 = (\pi_{ij})/(1-\pi_{ij})$ e não precisa ser estimada separadamente. Alguns softwares permitem a estimação de um fator de escala para a variância no menor nível, mas em geral, esse fator é definido como 1 (esse é o caso do SAS), isto é, supõe-se que os erros observados seguem exatamente o erro da distribuição teórica binomial.

Se o fator de escala for significativamente maior ou menor que 1, temos *overdispersion* ou *underdispersion*, respectivamente. Isso só pode ser estimado se o número de ensaios de bernoulli for maior que 1, o que não é o caso nesse trabalho. A ocorrência de *overdispersion* pode ser devido a valores extremos (*outliers*), omitir efeitos aleatórios importantes ou até mesmo uma quantidade pequena de grupos no segundo nível (por volta de 3, ver Wright,

1997). *underdispersion* pode ser devido a má especificação do modelo, como não incluir interações altamente significativas.

Por ser um modelo mais complicado, os métodos de estimação usados são todos numéricos. Incluir muitos parâmetros pode certamente levar a problemas de convergência do algoritmo para estimação. Geralmente os métodos utilizados são modificações do método da máxima verossimilhança, como *marginal quasi-likelihood* e *penalized quasi-likelihood*. Algumas vezes também se usa o método de máxima verossimilhança com algumas aproximações numéricas mais avançadas.

Outra questão que surge é a do coeficiente de correlação intra-classe. Com o fator de escala igual a 1, a variância que devemos usar é dada por $\pi^2/3 \approx$ 3,29, onde $\pi \approx$ 3,14 (Evans, Hastings e Peacock, 2000). Dessa forma, temos:

$$\rho = \frac{\sigma_G^2}{\sigma_G^2 + 3,29}$$

Nesse caso, ρ tem a mesma interpretação que na regressão linear multinível.

2.4 Modelos com Efeitos Mistos

Como já foi mencionado anteriormente, o modelo multinível tem efeitos fixos e mais de um efeito aleatório. Nesse caso, dizemos que este é um modelo misto, ou modelo com efeitos mistos. Nesta seção será definida a notação matricial desse modelo, com enfoque para regressão multinível, com 2 níveis. Além disso, o processo de estimação será apresentado no caso linear multinível (Modelo Linear Misto) e logístico multinível (Modelo Linear Generalizado Misto).

Estimação para esse tipo de modelo é um assunto muito extenso, pois muitos parâmetros precisam ser estimados e muitos métodos podem ser utilizados. Para o modelo linear misto, será considerado o método de máxima verossimi-

lhança restrita, que devido ao avanço dos computadores e as propriedades dos estimadores de máxima verossimilhança, produz bons resultados, além de que em muitas aplicações os grupos tem tamanhos diferentes tornando este um método melhor que a ANOVA, contanto que os pressupostos sejam verificados no diagnóstico.

2.4.1 Modelo Aleatório

O modelo aleatório foi brevemente comentado na seção anterior como um caso particular de modelo de componentes da variância. Como não apresenta variáveis explicativas, também é conhecido como modelo nulo, cuja utilidade no caso multinível é estimar as variâncias para calcular o coeficiente de correlação intraclasse. Vale observar que este ainda não é um modelo misto, pois o único efeito fixo é o intercepto.

Em 2.26, supomos que $G_i \sim$ N(0,σ^2_G), ou seja, tem média igual a zero. Isso parece ser questionável, mas pode ser visto como uma reparametrização, considerando um intercepto aleatório $\mu_i = \mu + G_i \sim$ N(μ,σ^2_G). Assim:

$$y_{ij} = \mu_i + e_{ij} = \mu + G_i + e_{ij}$$

Dessa forma, com a independência entre G_i e e_{ij}, tem-se que a covariância entre indivíduos de grupos diferentes é (i $\neq$ k):

$$\begin{aligned} Cov(y_{ij}, y_{kj}) &= E(y_{ij}y_{kj}) - E(y_{ij})E(y_{kj}) \\ &= E([\mu + G_i + e_{ij}][\mu + G_k + e_{kj}]) - \mu^2 \\ &= \mu^2 - \mu^2 \\ &= 0 \end{aligned}$$

E para indivíduos dentro do mesmo grupo (j $\neq$ k):

$$
\begin{aligned}
Cov(y_{ij}, y_{ik}) &= E(y_{ij}y_{ik}) - E(y_{ij})E(y_{ik}) \\
&= E([\mu + G_i + e_{ij}][\mu + G_i + e_{ik}]) - \mu^2 \\
&= \mu^2 + E(G_i^2) - \mu^2 \\
&= \sigma_G^2
\end{aligned}
$$

Logo obtém-se 2.27, o coeficiente de correlação intra-classe:

$$
\rho = \frac{Cov(y_{ij}, y_{ik})}{\sqrt{Var(y_{ij})Var(y_{ik})}} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}
$$

A notação matricial do modelo é dada por:

$$
\mathbf{Y} = \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\xi} \tag{2.34}
$$

- $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}\sigma_G^2)$; $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}\sigma_E^2)$; Independentes entre si.
- Var(**Y**) = $\mathbf{ZZ}'\sigma_G^2$ + $\mathbf{I}\sigma_E^2 = \mathbf{V}$
- **Y** $\sim$ N($\mathbf{1}\mu$,**V**)

Por exemplo, considerando I = 2 grupos com n_i observações cada, 2 para o grupo 1 e 3 para o grupo 2, i = 1, 2 e j = 1, ..., n_i, temos:

$$
\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{23} \end{bmatrix}
$$

Com a matriz de variância e covariância de **Y** dada por:

$$\mathbf{V} = \begin{bmatrix} \sigma_G^2 + \sigma_E^2 & \sigma_G^2 & 0 & 0 & 0 \\ \sigma_G^2 & \sigma_G^2 + \sigma_E^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_G^2 + \sigma_E^2 & \sigma_G^2 & \sigma_G^2 \\ 0 & 0 & \sigma_G^2 & \sigma_G^2 + \sigma_E^2 & \sigma_G^2 \\ 0 & 0 & \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma_E^2 \end{bmatrix}$$

Com a suposição de normalidade para **Y**, pode-se usar o método de máxima verossimilhança para estimar os parâmetros do modelo. A função de verossimilhança é dada por:

$$L(\mu, \mathbf{V}|\mathbf{Y}) = \frac{exp\left(-\frac{1}{2}(\mathbf{Y}\text{-}\mathbf{1}\mu)'\mathbf{V}^{\text{-}\mathbf{1}}(\mathbf{Y}\text{-}\mathbf{1}\mu)\right)}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \tag{2.35}$$

E a log-verossimilhança é:

$$l(\mu, \mathbf{V}|\mathbf{Y}) = l = -\frac{1}{2}(\mathbf{Y}\text{-}\mathbf{1}\mu)'\mathbf{V}^{\text{-}\mathbf{1}}(\mathbf{Y}\text{-}\mathbf{1}\mu) - \frac{n}{2}log(2\pi) - \frac{1}{2}log(|\mathbf{V}|) \tag{2.36}$$

Derivando 2.36 e igualando a zero, obtemos o sistema de equações normais, cuja solução resulta nos estimadores de máxima verossimilhança. As propriedades de derivadas de matrizes podem ser vistas em Searle et al, 2006. As derivadas são:

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= -\frac{1}{2}\frac{\partial((\mathbf{Y\text{-}1}\mu)'\mathbf{V^{-1}}(\mathbf{Y}-\mathbf{1}\mu))}{\partial \mu} \\
&= -\frac{1}{2}\frac{\partial(\mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y}-\mathbf{Y}'\mathbf{V}^{-1}\mathbf{1}\mu-\mathbf{1}'\mu\mathbf{V}^{-1}\mathbf{Y}+\mathbf{1}'\mu\mathbf{V}^{-1}\mathbf{1}\mu)}{\partial \mu} \\
&= -\frac{1}{2}(-\mathbf{Y}'\mathbf{V}^{-1}\mathbf{1}-\mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}+2\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}\mu) \\
&= -\frac{1}{2}(-2\mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}+2\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}\mu) \\
&= \mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}-\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}\mu
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma^2_G} &= -\frac{1}{2}\frac{\partial log(|\mathbf{V}|)}{\partial \sigma^2_G} - \frac{1}{2}\frac{\partial((\mathbf{Y\text{-}1}\mu)'\mathbf{V^{-1}}(\mathbf{Y\text{-}1}\mu))}{\partial \sigma^2_G} \\
&= -\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') - \frac{1}{2}(\mathbf{Y\text{-}1}\mu)'\frac{\partial \mathbf{V}^{-1}}{\partial \sigma^2_G}(\mathbf{Y\text{-}1}\mu) \\
&= -\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') + \frac{1}{2}(\mathbf{Y\text{-}1}\mu)'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y\text{-}1}\mu)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma^2_E} &= -\frac{1}{2}tr(\mathbf{V}^{-1}\mathbf{I}\mathbf{I}') + \frac{1}{2}(\mathbf{Y\text{-}1}\mu)'\mathbf{V}^{-1}\mathbf{I}\mathbf{I}'\mathbf{V}^{-1}(\mathbf{Y\text{-}1}\mu) \\
&= -\frac{1}{2}tr(\mathbf{V}^{-1}) + \frac{1}{2}(\mathbf{Y\text{-}1}\mu)'\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{Y\text{-}1}\mu)
\end{aligned}
$$

Obtém-se então o seguinte sistema de equações normais:

$$
\begin{cases}
\mathbf{1}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{1}'\mathbf{V}^{-1}\mathbf{1}\mu \\
(\mathbf{Y\text{-}1}\mu)'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y\text{-}1}\mu) = tr(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') \\
(\mathbf{Y\text{-}1}\mu)'\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{Y\text{-}1}\mu) = tr(\mathbf{V}^{-1})
\end{cases} \tag{2.37}
$$

As variâncias assintóticas para o vetor de parâmetros $\boldsymbol{\theta}$ são obtidas por Var($\boldsymbol{\theta}$) $\cong [\mathbf{I}(\boldsymbol{\theta})]^{-1}$ onde $\mathbf{I}(\boldsymbol{\theta})$ é a matriz de informação de Fisher. Os cálculos dessa matriz para esse modelo são extensos mas não muito complicados e podem ser vistos em Searle et al, 2006.

O método de estimação de máxima verossimilhança restrita é atualmente

usado da seguinte forma (veja Searle et al, 2006):

1. Multiplicar 2.34 por uma matriz **P**, tal que **P1** $= \mathbf{0}$, resultando em :

$$\mathbf{PY} = \mathbf{PZ}\boldsymbol{\alpha} + \mathbf{P}\boldsymbol{\xi} \tag{2.38}$$

   Assim, elimina-se o efeito fixo μ, restringindo o espaço dos parâmetros aos componentes da variância. As propriedades de **P** podem ser vistas em Searle et al, 2006.

2. Aplicar o método de máxima verossimilhança em **PY** $\sim$ N(**0**,**PVP**$'$)

3. substituir as estimativas dos componentes da variância em:

$$\hat{\mu} = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}$$

O sistema de equações normais é obtido da mesma forma como feito anteriormente.

2.4.2 Modelos Mistos para Regressão Linear Hierárquica

A notação usual de um modelo misto é:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\xi} \tag{2.39}$$

Onde **X**β é a matriz de efeitos fixos, **Z**α é a matriz de efeitos aleatórios e ξ é a matriz dos erros, também aleatórios.

O modelo de regressão linear multinível se encaixa nessa notação. Primeiramente considerando o modelo com todos os efeitos fixos, mas sem coeficientes de regressão aleatórios, dado por:

$$y_{ij} = \mu + \sum_{p=1}^{P} \beta_p x_{pij} + \sum_{q=1}^{Q} \gamma_q w_{qi} + \sum_{p=1}^{P}\sum_{q=1}^{Q} \theta_{pq} x_{pij} w_{qj} + G_i + e_{ij} \tag{2.40}$$

Tem-se que 2.40 tem seus efeitos fixos denotados matricialmente por:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{w} & \mathbf{k} \end{bmatrix} \begin{bmatrix} \mu \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \\ \boldsymbol{\theta} \end{bmatrix} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\xi} \tag{2.41}$$

Assim, em 2.41 temos a mesma notação matricial do modelo de componentes da variância para os efeitos aleatórios. Para os efeitos fixos, todos são juntados na mesma matriz. Vale observar que a matriz **X** tem posto completo, logo a solução de máxima verossimilhança para $\boldsymbol{\beta}$ é unica.

Novamente considerando o exemplo em que se tem I $= 2$ grupos com n_i observações cada, 2 para o grupo 1 e 3 para o grupo 2, i $= 1, 2$ e j $= 1, ..., n_i$, com duas variáveis explicativas x_1 e x_2 no menor nível e uma variável w_1 no maior nível, e considerando interação entre níveis com as variáveis x_1 e w_1, temos que p $=$ q $= 1$ e a notação matricial baseada em 2.40 e 2.41 é:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & x_{111} & x_{211} & w_{11} & x_{111}w_{11} \\ 1 & x_{112} & x_{212} & w_{11} & x_{111}w_{11} \\ 1 & x_{121} & x_{221} & w_{12} & x_{111}w_{12} \\ 1 & x_{122} & x_{222} & w_{12} & x_{111}w_{12} \\ 1 & x_{123} & x_{223} & w_{12} & x_{111}w_{12} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \theta_{11} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{23} \end{bmatrix}$$

Para estimar os parâmetros utilizando máxima verossimilhança, basta seguir os mesmos passos feitos no modelo de componentes da variância, chegando ao seguinte sistema de equações normais:

$$\begin{cases} \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \\ (\mathbf{Y}\text{-}\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y}\text{-}\mathbf{X}\boldsymbol{\beta}) = tr(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}') \\ (\mathbf{Y}\text{-}\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{Y}\text{-}\mathbf{X}\boldsymbol{\beta}) = tr(\mathbf{V}^{-1}) \end{cases} \tag{2.42}$$

O método de máxima verossimilhança restrita também é análogo. Multiplica-se 2.41 por uma matriz **P**, tal que **PX** $=$ **0**, e aplica-se o método de máxima verossimilhança em **PY** $\sim$ N(**0**,**PVP**′). Para estimar $\boldsymbol{\beta}$, basta substituir $\hat{\mathbf{V}}^{-1}$ em $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$.

Se os coeficientes de regressão no menor nível forem aleatórios, o processo de estimação é mais complicado, envolvendo também a derivada da função de verossimilhança em relação a variância desse feito.

Outra posssível aplicação de modelos mistos para regressão linear hierárquica é o caso em que se tem 3 níveis na hierarquia, mas apenas 2 desses níveis são considerados aleatórios. A questão de um efeito ser aleatório ou fixo depende muito do contexto dos dados ou do interesse do pesquisador, como por exemplo, se o interesse for na variabilidade dentro de cada grupo do terceiro nível, isto é, analisar como são as inclinações de cada variável explicativa dentro de cada um desses grupos, pode-se considerar o terceiro nível como fixo.

Considere que uma população hierárquica tenha 3 níveis, sendo o maior nível composto por 2 grupos, i $=$ 1, 2, denotados por S_i, sendo considerado efeito fixo. O segundo nível também com 2 grupos, j $=$ 1, 2, $\forall$ i, sendo considerado aleatório. No menor nível temos $k_j = 2$ para j $=$ 1 e $k_j = 3$ para j $=$ 2, sendo também considerado aleatório. Considerando apenas uma variável explicativa x no menor nível, tal que as inclinações de x sejam diferentes em cada S_i, porém fixas, o modelo é dado por:

$$y_{ijk} = \mu + \beta_i x_{ijk} + s_i + G_j + e_{ijk} \tag{2.43}$$

Sua notação matricial é a mesma de 2.41 para a matriz de efeitos aleatórios e o erro, mas para a matriz de efeitos fixos temos:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1} & x_1 & x_1 & \mathbf{s} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ s_1 \\ s_2 \end{bmatrix} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\xi} \tag{2.44}$$

Resultando em:

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{223} \end{bmatrix} = \begin{bmatrix} 1 & x_{111} & 0 & 1 & 0 \\ 1 & x_{112} & 0 & 1 & 0 \\ 1 & x_{121} & 0 & 1 & 0 \\ 1 & x_{122} & 0 & 1 & 0 \\ 1 & x_{123} & 0 & 1 & 0 \\ 1 & 0 & x_{211} & 0 & 1 \\ 1 & 0 & x_{212} & 0 & 1 \\ 1 & 0 & x_{221} & 0 & 1 \\ 1 & 0 & x_{222} & 0 & 1 \\ 1 & 0 & x_{223} & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{123} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \\ e_{223} \end{bmatrix}$$

Vale observar que a matriz **X** tem posto incompleto, devido a matriz **s**, pois somando a terceira e quarta coluna de **X** obtemos a primeira coluna. Porém, pode-se mostrar que μ e β_i tem solução única, enquanto que s_i tem infinitas soluções.

2.4.3 Modelos Lineares Generalizados Mistos para Regressão Logística Hierárquica

Assim como a classe de modelos lineares generalizados é a forma geral para modelos como regressão linear e regressão logística, a classe de modelos lineares generalizados mistos segue o mesmo propósito, só que para modelos mistos, ou seja, MLGM tem em seu componente sistemático efeitos aleatórios e variáveis resposta de outros níveis. Nesta seção, será definido o modelo de regressão logística multinível como um modelo linear generalizado misto, que será usado no banco de dados desse trabalho.

Usando a notação matricial da seção anterior, temos que o modelo é dado por:

$$logito(\pi_{ij}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} \tag{2.45}$$

Onde tem-se k efeitos aleatórios e $\boldsymbol{\alpha} \sim$ N(0,$\mathbf{D}_*$), com $\mathbf{D}_* = \sigma_G^2\mathbf{D}$, para facilitar os cálculos. O método de máxima verossimilhança leva a uma integral k-dimensional. Considere que $\mathbf{D}_-$ = $\mathbf{D}_*^{-1}$ é a matriz de precisão. A função de log-verossimilhança é dada por:

$$l(\boldsymbol{\beta}, \mathbf{D}_-) = -\frac{Ik}{2}log(2\pi) + \frac{I}{2}log(|\mathbf{D}_-|) + \boldsymbol{\beta}\mathbf{r} + \sum_{i=1}^{I} log\left(\int_{\Re^k} e^{h_i(\boldsymbol{\beta},\mathbf{u})}d\mathbf{u}\right) \tag{2.46}$$

Onde tem-se I grupos com n_i observações cada, e β tem m parâmetros. Além disso, temos em 2.46:

$$\mathbf{r} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\mathbf{Y}_{ij}\mathbf{X}_{ij},$$

$$h_i(\boldsymbol{\beta}, \mathbf{u}) = \mathbf{k}_i'\mathbf{u} - \frac{1}{2}\mathbf{u}'\mathbf{D}_-\mathbf{u} - \sum_{j=1}^{n_i} log(1 + e^{\boldsymbol{\beta}\mathbf{X}_{ij}+\mathbf{u}'\mathbf{Z}_{ij}}),$$

$$\mathbf{k}_i = \sum_{j=1}^{n_i} \mathbf{Y}_{ij}\mathbf{Z}_{ij}$$

Observar que **r** é um vetor m x 1 e $\mathbf{k}_i$ é um vetor k x 1. As derivadas de primeira ordem são:

$$\begin{cases} \dfrac{\partial l(\boldsymbol{\beta}, \mathbf{D}_-)}{\partial \boldsymbol{\beta}} = \mathbf{r} - \displaystyle\sum_{i=1}^{I} \frac{\mathbf{l}_{i3}}{I_{i1}} \\ \dfrac{\partial l(\boldsymbol{\beta}, \mathbf{D}_-)}{\partial \mathbf{D}_-} = \dfrac{1}{2}\left(I\mathbf{D}_-^{-1} - \displaystyle\sum_{i=1}^{I} \frac{\mathbf{l}_{i2}}{I_{i1}} \right) \end{cases} \tag{2.47}$$

Onde em 2.47 temos:

$$I_{i1} = \int_{\Re^k} e^{h_i(\boldsymbol{\beta},\mathbf{u})} d\mathbf{u}$$

$$\mathbf{l}_{i2} = \int_{\Re^k} \mathbf{u}\mathbf{u}' e^{h_i(\boldsymbol{\beta},\mathbf{u})} d\mathbf{u}$$

$$\mathbf{l}_{i3} = \int_{\Re^k} \left[X_{ij} \frac{e^{\boldsymbol{\beta}\mathbf{X}_{ij} + \mathbf{u}'\mathbf{Z}_{ij}}}{1 + e^{\boldsymbol{\beta}\mathbf{X}_{ij} + \mathbf{u}'\mathbf{Z}_{ij}}} e^{h_i(\boldsymbol{\beta},\mathbf{u})} \right] d\mathbf{u}$$

Tem-se que $\dfrac{\partial l(\boldsymbol{\beta}, \mathbf{D}_-)}{\partial \mathbf{D}_-} = \mathbf{0}$ é fácil de computar devido a parametrização escolhida. Como as derivadas de primeira ordem envolvem muitas integrais, alguns métodos de aproximação são muito úteis, já que muitas vezes essas integrais não convergem. O problema da estimação é ainda maior quando se tem muitos efeitos aleatórios, pois a integral envolvida é k-Dimensional. Demidenko, 2004 apresenta vários métodos de maximização. Em particular, o método de Laplace foi utilizado nesse trabalho.

Capítulo 3

Aplicação

Neste capítulo será apresentada uma típica aplicação de regressão hierárquica, que consiste de um estudo em que os interesses são os fatores que influenciam na aprovação do aluno.

3.1 Introdução

A crescente disponibilidade de dados, de diversas formas, tem aumentado não só a demanda por análises estatísticas das mais simples até as mais sofisticadas, mas também a demanda por profissionais com conhecimento no mínimo básico de estatística.Sendo assim, a ciência Estatística evoluiu muito nas últimas décadas devido à essa demanda e ao avanço dos computadores, possibilitando análises cada vez mais complexas e contribuindo com várias pesquisas nas mais diversas áreas do conhecimento. Dessa forma, diversos cursos de graduação e pós-graduação têm disciplinas da ciência Estatística como obrigatórias em seus currículos.

Por esse motivo, as disciplinas ofertadas pelo departamento de Estatística da UnB para outros cursos, também chamadas de disciplinas de serviço, são de extrema importância para a Universidade. Atualmente são ofertadas 3 disc-

plinas de serviço: Bioestatística, que é obrigatória para os cursos Agronomia, Engenharia Florestal e Medicina Veterinária; Estatística Aplicada, obrigatória para Administração, Arquivologia, Biblioteconomia, Ciências Ambientais, Ciências Contábeis, Ciências Sociais, Ciência Política, Geografia, Gestão de Agronegócios, Psicologia e Relações Internacionais; e por fim Probabilidade e Estatística, obrigatória para Ciências Econômicas, Ciência da Computação, Engenharia Civil, Engenharia Elétrica, Engenharia Mecânica, Engenharia de Computação, Engenharia de Redes de Comunicação, Engenharia de Produção e Engenharia Química.

Uma preocupação dos professores do departamento de Estatística é quais fatores influenciam o rendimento dos alunos que cursam as disciplinas acima citadas. Este trabalho apresenta uma análise desses fatores e seus efeitos em cada uma dessas disciplinas como um exemplo de aplicação do modelo estudado.

Visto que a população considerada tem estrutura hierárquica com 3 níveis, isto é, alunos agrupados dentro de turmas, que estão agrupadas dentro de disciplinas, um modelo adequado é a regressão multinível.

3.2 Metodologia

Nesse estudo foram utilizados os dados referentes aos alunos que cusaram as disciplinas de serviço do curso de Estatística da UnB no período de 2004 a 2008, que foram obtidos pelo SIGRA, Sistema de Informação Acadêmica de Graduação, da UnB. O foco do trabalho é o ano de 2008, mas os anos de 2004 a 2007 foram utilizados para apresentar um panorama geral do rendimento dos alunos nesse período. Vale observar que trata-se de um estudo com dados observacionais, visto que nenhum tipo de amostragem foi utilizado na coleta dos dados.

Inicialmente foi feita uma análise descritiva dos dados, apresentando as

principais características dos alunos e das turmas. Após essa etapa, uma análise descritiva bivariada foi feita para identificar quais variáveis que mais influenciam no desempenho dos alunos, sendo a variável resposta do estudo binária: aprovado ou reprovado. Para as variáveis qualitativas, fez-se uso do teste qui-quadrado de associação, com correção de continuidade quando necessário. Para as variáveis quantitativas ajustou-se um modelo de regressão logística simples, verificando a significância da variável considerada.

As variáveis consideradas significativas na análise bivariada foram consideradas na modelagem. O ajuste do modelo foi feito inicialmente sem variáveis explicativas para calcular o coeficiente de correlação intraclasse e testar se as variâncias em diferentes turmas são homogêneas. Após essa etapa, as variáveis explicativas foram colocadas uma de cada vez e verificou-se a significância e o BIC de cada modelo, sendo o modelo escolhido nesse primeiro passo o que apresentar o coeficiente significativo e o menor BIC. No segundo passo, novas variáveis foram adicionadas até que nenhuma outra fosse significativa, chegando a um ou vários candidatos a modelo final.

Por fim foi feito o diagnóstico dos candidatos a modelo final, verificando os pressupostos e a qualidade do ajuste. Passado o diagnóstico, as inferências de interesse foram feitas. Visto que os dados tem estrutura hierárquica e a variável resposta é binária, o modelo utilizado foi o de regressão logística multinível, caso particular de um modelo linear generalizado misto.

Os métodos de estimação utilizados nos modelos foi de *Residual Pseudo-Likelihood*. Em alguns casos utilizou-se o método de Máxima Verossimilhança com a aproximação de Laplace.

O nível disciplina foi considerado separadamente, de modo que um modelo foi feito para cada disciplina, visto que estas tinham diferentes variáveis explicativas importantes. Ambos os softwares SAS e R foram usados para analisar os dados, sendo que a modelagem foi feita apenas com o SAS.

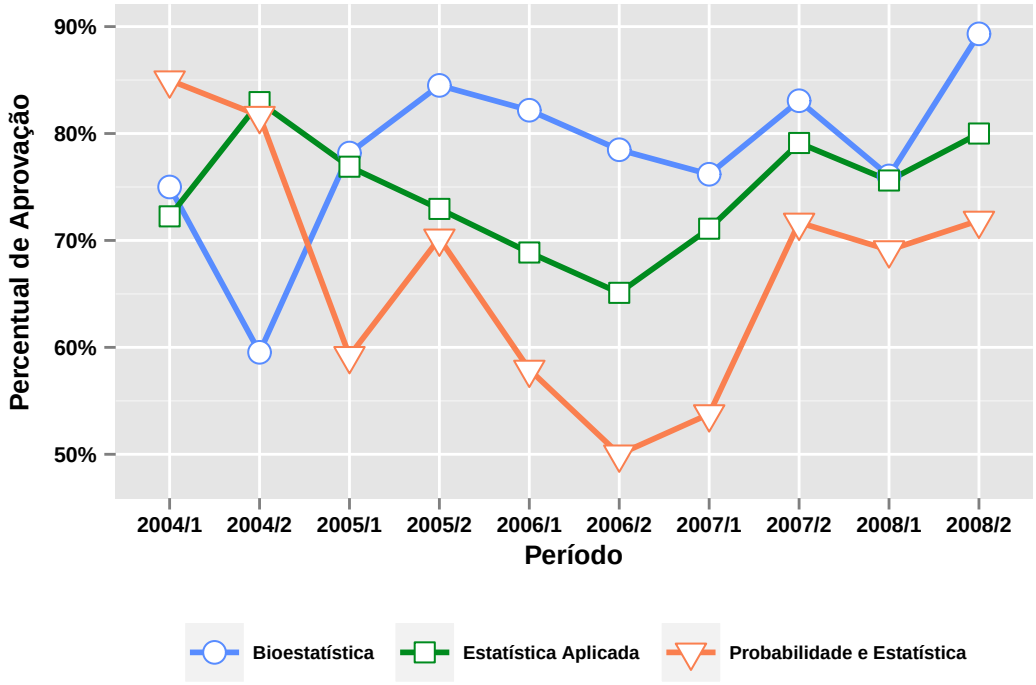
Capítulo 4

Análise Descritiva

4.1 Panorama Geral

O ano de interesse é o ano de 2008, mas, para apresentar a evolução histórica do percentual de aprovações, foi feito um gráfico do percentual de aprovação por disciplina para os períodos do primeiro semestre de 2004 ao segundo semestre de 2008. Na figura 4.1 tem-se que a disciplina Bioestatística que no início do período considerado estava com os menores percentuais de aprovação, passou a ter os maiores a partir de 2005/2, que é uma situação contrária a disciplina Probabilidade e Estatística, que começou com os maiores e passsou a ter os menores percentuais. Estatística Aplicada apresentou percentuais relativamente estáveis.

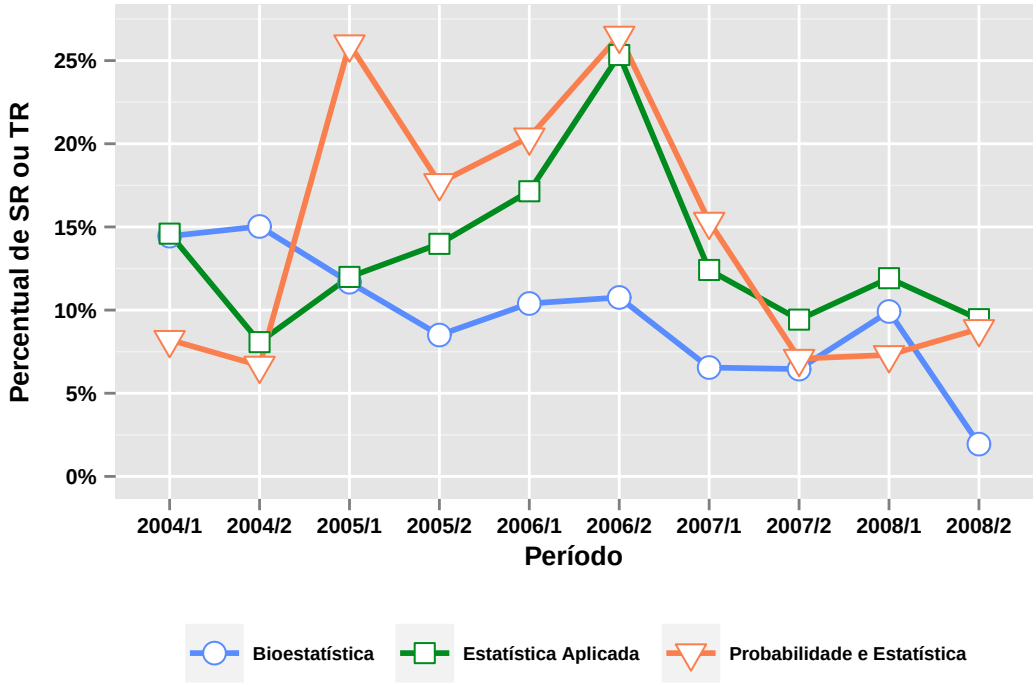
Figura 4.1: Percentual de Aprovação por Disciplina



PE	85.0%	81.7%	59.2%	70.2%	57.9%	50.0%	53.8%	71.7%	69.1%	71.9%
EA	72.3%	82.9%	76.9%	72.9%	68.9%	65.1%	71.1%	79.1%	75.6%	80.0%
BIO	75.0%	59.5%	78.2%	84.5%	82.2%	78.5%	76.2%	83.1%	76.0%	89.3%

Além disso, nesse período, a figura 4.2 apresenta o percentual de SR e TR. As disciplinas Probabilidade e Estatística Aplicada apresentam comportamento de altos e baixos semelhante, chegando a aproximadamente 25% em alguns períodos, enquanto que Bioestatística apresentou uma redução desse percentual e na maior parte dos períodos o menor percentual, com apenas 1,94% de SR ou TR no período de 2008/2.

Figura 4.2: Percentual de SR ou TR por Disciplina



PE	8.2%	6.7%	26.0%	17.6%	20.4%	26.5%	15.3%	7.1%	7.3%	8.9%
EA	14.6%	8.1%	12.0%	14.0%	17.1%	25.3%	12.4%	9.4%	11.9%	9.5%
BIO	14.4%	15.0%	11.7%	8.5%	10.4%	10.8%	6.5%	6.5%	9.9%	1.9%

Como as menções SR e TR indicam que o aluno não foi a maior parte das aulas ou desistiu da disciplina, esses serão desconsiderados da modelagem. Dessa forma, a figura 4.3 apresenta os percentuais de aprovação no período de análise, desconsiderando SR ou TR.

Figura 4.3: Percentual de Aprovação por Disciplina sem SR e TR

PE	92.6%	87.5%	80.0%	85.2%	72.8%	68.0%	63.5%	77.1%	74.5%	78.9%
EA	84.6%	90.2%	87.4%	84.8%	83.1%	87.2%	81.2%	87.3%	85.8%	88.4%
BIO	87.7%	70.1%	88.5%	92.3%	91.7%	87.9%	81.5%	88.8%	84.4%	91.1%

Nesse caso, a disciplina Probabilidade e Estatística apresenta percentuais visivelmente menores na maior parte dos períodos, inclusive no ano de 2008, enquanto que Estatística Aplicada e Bioestatística estão praticamente iguais.

No ano de 2008, sem os alunos que tiveram menção SR ou TR, Bioestatística teve 210 alunos, Estatística Aplicada 971 e Probabilidade e Estatística teve 628 alunos. No total, tem-se 1809 alunos.

4.2 Nível Aluno

Antes de apresentar as características dos alunos que influenciam em sua aprovação, a tabela abaixo mostra qual o perfil dos alunos em cada disciplina nos dois períodos do ano de 2008.

Tabela 4.1: Características dos Estudantes

Variáveis		Disciplina						
		Bio		EA		PE		
		2008/1	2008/2	2008/1	2008/2	2008/1	2008/2	Geral
Sexo	Feminino	60,5%	51,5%	48,4%	48,0%	16,1%	11,7%	37,2%
	Masculino	39,5%	48,5%	51,6%	52,0%	83,9%	88,3%	62,8%
País de Nascimento	Brasil	99,1%	100%	98,1%	97,6%	98,5%	99,3%	98,4%
	Exterior	0,9%	-	1,9%	2,4%	1,5%	0,7%	1,6%
Naturalidade(1)	Distrito Federal	66,7%	75,2%	66,2%	67,5%	64,3%	64,2%	66,4%
	Goiás	8,3%	8,9%	6,4%	7,4%	10,8%	11,1%	8,5%
	Minas Gerais	9,3%	3,0%	5,6%	5,1%	2,8%	5,7%	5,0%
	Rio de Janeiro	2,8%	5,0%	2,2%	4,5%	4,0%	4,1%	3,6%
	São Paulo	3,7%	-	3,9%	2,0%	4,9%	3,0%	3,2%
	Outras	9,2%	16,8%	15,7%	13,5%	13,2%	11,9%	13,3%
UF de residência	Distrito Federal	91,7%	91,1%	97,0%	96,4%	93,0%	94,6%	95,1%
	Outras	8,3%	8,9%	3,0%	3,6%	7,0%	5,6%	4,9%
RA de residência	Brasília	60,6%	53,4%	61,6%	62,4%	58,5%	50,6%	58,9%
	Taguatinga	11,9%	15,8%	8,2%	9,4%	9,4%	12,8%	10,2%
	Sobradinho	6,4%	3,0%	3,4%	3,4%	2,7%	4,3%	3,6%
	Guará	4,6%	5,0%	2,7%	3,4%	2,7%	4,4%	3,4%
	Outras	16,5%	22,8%	24,1%	21,4%	26,7%	27,9%	23,9%
Local(2)	DF Alta Renda	62,3%	56,4%	62,3%	63,4%	63,9%	57,7%	61,9%
	DF Média Renda	27,5%	28,7%	23,7%	23,7%	22,4%	30,5%	25,1%
	DF Baixa Renda	3,7%	6,9%	11,7%	10,0%	7,2%	6,0%	8,7%
	GO Entorno	1,8%	4,0%	1,1%	1,2%	1,5%	1,3%	1,4%
	Outros	4,7%	4,0%	1,2%	1,7%	5,0%	4,5%	2,9%

Notas: - Dado numérico igual a zero não resultante de arredondamento.

(1) Naturalidade: considera apenas os brasileiros.

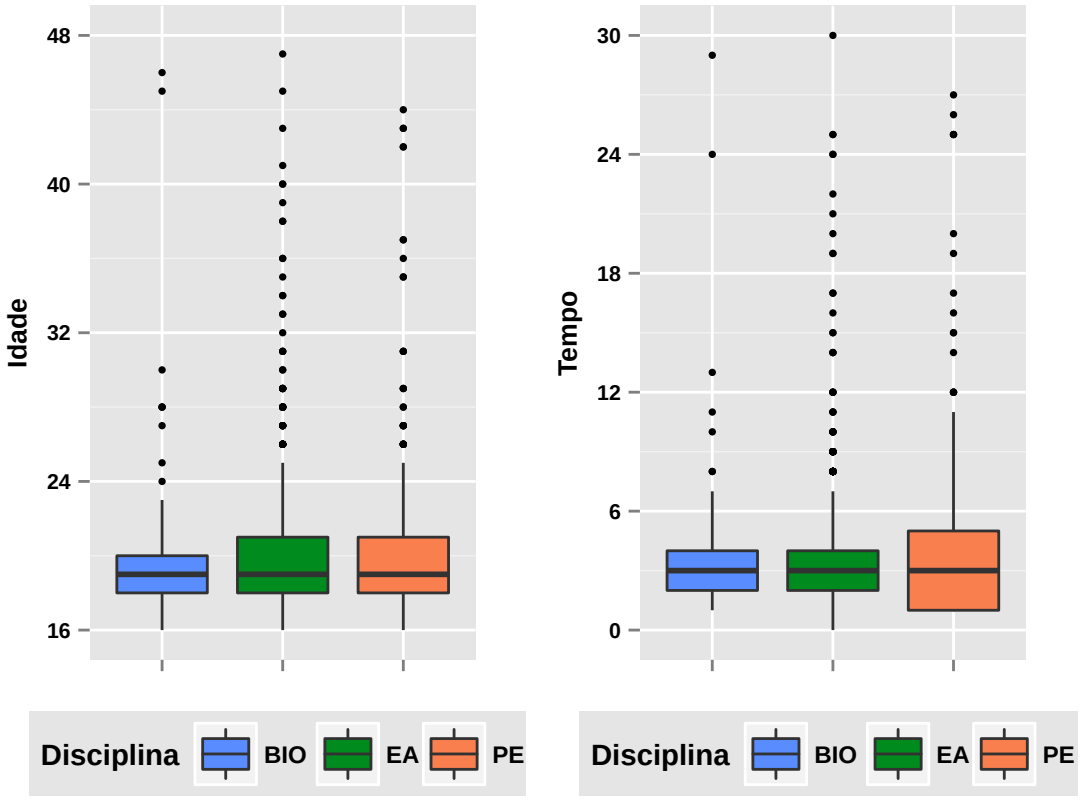
(2) As regiões administrativas Sudoeste/Octogonal, Brasília, Lago Sul, Lago Norte e Park Way foram consideradas de alta renda, já as RA's Águas Claras, Candangolândia, Cruzeiro, Cruzeiro Velho, Gama, Guará, Núcleo Bandeirante, Riacho Fundo, São Sebastião, Sobradinho, Sobradinho II, Taguatinga, Vicente Pires de média renda, Brazlândia, Ceilândia, Paranoá, Planaltina, Recanto das Emas, Riacho Fundo II, Samambaia, Santa Maria, Valparaíso e Valparaíso II de baixa renda e por fim Águas Lindas, Cidade Ocidental, Formosa, Luziânia, Novo Gama, Santo Antônio do Descoberto e Valparaíso de Goiás do entorno.

Na disciplina Bioestatística, pode-se perceber uma maioria do sexo feminino, embora no segundo período de 2008 não tenha tanta diferença. Para a disciplina estatística aplicada, os percentuais de ambos sexos estão equilibrados. Por outro lado, em Probabilidade e Estatística, em torno de 85% dos estudantes é do sexo masculino, indicando uma minoria do sexo feminino. Em todas as disciplinas, a grande maioria dos estudantes nasceram no Brasil, tem naturalidade no DF e residem no DF.

Mais de 50% dos alunos residem em Brasília, seguido de 10% em Taguatinga e por volta de 3,5% no Sobradinho e no Guará, não aparentando haver muita diferença entre disciplinas. Da mesma forma, mais de 50% dos alunos em todas as disciplinas são de alta renda, seguido de em torno de 25% de média renda. Já para os de baixa renda, tem-se uma leve diferença entre disciplinas, com estatística aplicada liderando com mais de 10% dos alunos de baixa renda. Do entorno e de outras localidades tem-se em torno de 5% dos alunos.

Os boxplots na figura 4.4 apresentam como se distribuem as variáveis quantitativas idade e tempo desde a conclusão do ensino médio, em anos. Por esse gráfico pode-se ter uma idéia de que temos vários valores discrepantes, que podem ou não impactar na modelagem. Para todos as disciplinas a mediana da idade é bem próxima, com Bioestatística apresentando menor variabilidade. Analogamente para a variável tempo, a mediana é bem próxima nas 3 disciplinas, com uma variabilidade maior em probabilidade e estatística.

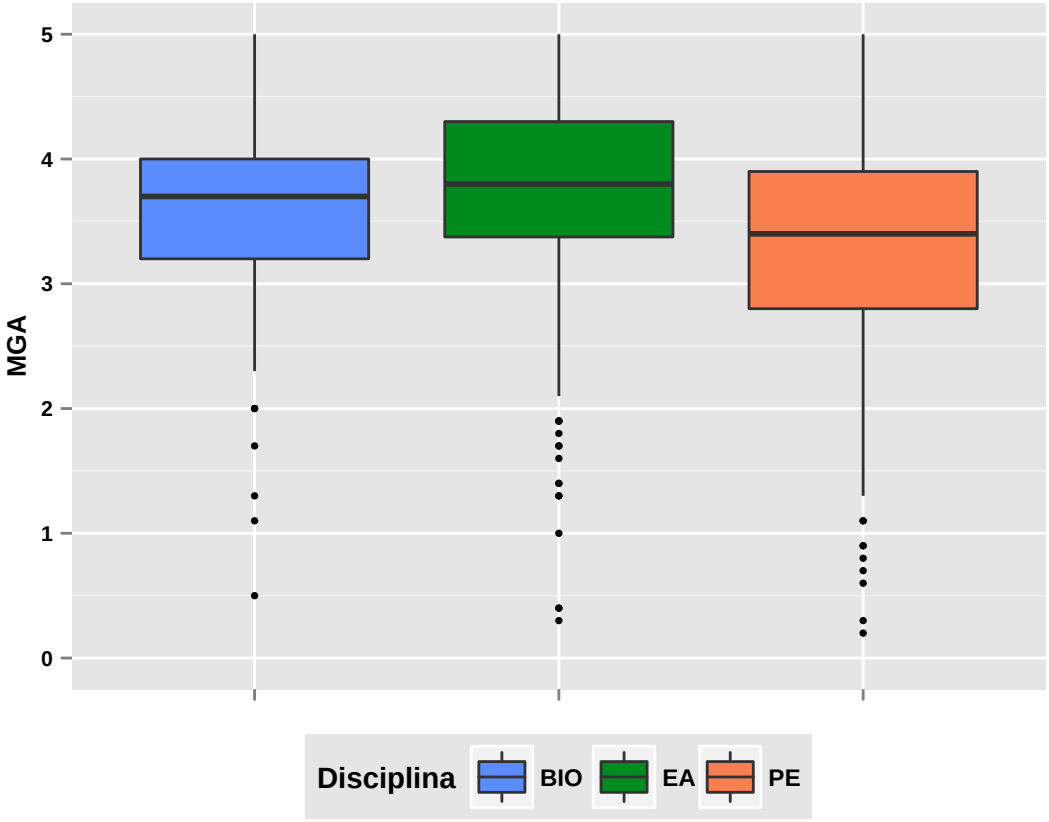
Figura 4.4: *Boxplot* das Variáveis Idade e Tempo desde a Conclusão do Ensino Médio



Espera-se que a variável idade e tempo desde a conclusão do ensino médio sejam correlacionadas, então o coeficiente de correlação linear foi calculado resultando em 97% para Bioestatística, 91% para Estatística Aplicada e 97% para Probabilidade e Estatística.

O boxplot na figura 4.5 se refere a variável MGA (Média Geral Acumulada), que é o índice de rendimento acadêmico do aluno no semestre em que ele cursou a disciplina em questão.

Figura 4.5: *Boxplot* da Variável Média Geral Acumulada



Pode-se também perceber varios valores discrepantes, alunos com MGA baixo. Além disso, a disciplina Probabilidade e Estatística apresenta maior variabilidade do que as outras disciplinas, com uma maior quantidade de alunos com valores pequenos de MGA.

As variáveis quantitativas discutidas anteriormente apresentam dados faltantes *(missings)*. Em Bioestatística tem-se 34 *missings* de 210 observações, Estatística Aplicada com 135 de 971 e Probabilidade e Estatística 93 de 628, totalizando 262 *missings* de 1809 observações.

A tabela 4.2 apresenta as características acadêmicas dos alunos.

Tabela 4.2: Características Acadêmicas dos Estudantes

		Disciplina						
		Bio		EA		PE		
	Variáveis	2008/1	2008/2	2008/1	2008/2	2008/1	2008/2	Geral
Ano de Ingresso	2004	7,3%	1,0%	2,3%	1,8%	8,5%	4,4%	3,9%
	2005	13,8%	5,0%	9,9%	7,4%	9,7%	12,4%	9,6%
	2006	11,0%	5,9%	19,0%	13,5%	12,7%	10,4%	13,7%
	2007	67,0%	61,4%	60,5%	19,5%	60,0%	24,1%	43,6%
	2008	0,9%	23,8%	4,9%	55,2%	3,9%	47,0%	26,3%
	Antes de 2004	-	2,9%	3,4%	2,6%	5,2%	1,7%	2,9%
Forma de Ingresso	Vestibular	73,4%	76,2%	80,3%	58,2%	84,6%	68,1%	72,4%
	PAS	24,8%	21,8%	9,7%	33,5%	11,6%	26,8%	21,0%
	TFO	-	1,0%	1,9%	3,4%	2,1%	4,4%	2,6%
	TFF	1,8%	1,0%	4,0%	1,8%	0,3%	0,3%	1,8%
	Outras	-	-	4,1%	3,1%	1,4%	0,4%	2,2%
Curso	Administração	-	-	23,3%	19,1%	-	-	11,3%
	Agronomia	20,2%	23,8%	-	0,6%	-	-	2,7%
	Arquivologia	-	-	9,1%	8,8%	-	-	4,8%
	Biblioteconomia	-	-	10,1%	10,4%	-	-	5,5%
	Ciência da Computação	-	-	-	-	10,3%	10,4%	3,6%
	Ciência Política	-	-	7,0%	7,6%	-	-	3,9%
	Ciências Biológicas	13,8%	2,0%	-	0,2%	0,3%	-	1,1%
	Ciências Contábeis	-	-	14,2%	15,5%	0,3%	-	8,0%
	Ciências Econômicas	-	-	1,1%	0,8%	2,1%	2,0%	1,2%
	Ciências Sociais	-	-	11,0%	6,8%	-	-	4,8%
	Computação	-	-	-	-	10,9%	10,7%	3,8%
	Engenharia Civil	-	-	-	-	17,2%	18,5%	6,2%
	Engenharia de Redes de Comunicação	-	-	-	-	8,8%	11,7%	3,5%
	Engenharia Elétrica	-	-	-	-	11,6%	12,1%	4,1%
	Engenharia Florestal	30,3%	35,6%	-	0,2%	-	-	3,9%
	Engenharia Mecânica	0,9%	2,0%	-	-	10,9%	12,1%	4,1%
	Engenharia Mecatrônica	-	-	-	-	7,9%	9,1%	2,9%
	Farmácia	13,8%	7,9%	-	-	-	-	1,3%
	Geografia	-	-	6,8%	6,8%	-	-	3,6%
	Matemática	0,9%	-	0,4%	1,8%	17,9%	11,4%	5,8%
	Medicina Veterinária	17,4%	26,7%	-	-	-	0,3%	2,6%
	Psicologia	-	-	2,7%	8,4%	-	-	3,0%
	Relações Internacionais	-	-	8,0%	9,0%	-	-	4,6%
	Serviço Social	-	-	2,3%	1,0%	-	-	0,9%
	Outros	2,7%	2,0%	4,0%	3,0%	1,6%	1,7%	2,8%
Modalidade	Obrigatória	69,7%	87,1%	93,6%	95,0%	90,1%	94,0%	91,7%
	Optativa	29,4%	12,9%	5,1%	4,2%	9,6%	6,0%	7,6%
	Módulo Livre	0,9%	-	1,3%	0,8%	0,3%	-	0,7%
Menção	II	8,3%	4,0%	5,1%	6,4%	17,3%	4,7%	7,7%
	MI	7,4%	4,9%	9,1%	5,2%	8,2%	16,4%	8,7%
	MM	44,9%	54,5%	48,2%	36,7%	40,6%	41,3%	42,7%
	MS	30,3%	32,7%	29,8%	34,7%	26,7%	31,5%	31,1%
	SS	9,1%	3,9%	7,8%	17,0%	7,2%	6,1%	9,8%
Cotas	Sim	14,8%	17,0%	16,1%	12,7%	16,1%	13,2%	14,7%
	Não	85,2%	83,0%	83,9%	87,3%	83,9%	86,8%	85,3%

A maioria dos alunos tiveram ingresso em 2008 e 2007, sendo que o percentual de estudantes no primeiro semestre de 2008 em relação ao ano de 2007, em todas as disciplinas. A forma de ingresso mais comun é o vestibular, visto que nessa época a UnB tinha 2 vestibulares no ano. PAS (programa de avaliação seriada) está em torno de 20% com grandes variações entre o primeiro e o segundo semestre de estatística aplicada e probabilidade e estatística.

Os cursos apresentados na tabela são aqueles em que pelo menos uma das disciplinas consideradas é obrigatória. Pode-se perceber que a maioria dos estudantes de Bioestatística são da engenharia florestal e agronomia, para Estatística Aplicada temos administração ciências contábeis e biblioteconomia. Para Probabilidade e Estatística temos uma boa distribuição entre cursos, mas com maior percentual em engenharia civil. Em geral, a maioria dos alunos são de administração, com pouco mais de 11%.

Quanto a modalidade, os alunos cursam, em maior parte, as disciplinas de serviço de estatística como obrigatórias. A disciplina Bioestatística apresenta um maior percentual de alunos que cursaram como optativa, mas em todos as disciplinas o percentual de módulo livre é praticamente nulo. Em torno de 70% das menções foram de MM e MS, com baixos percentuais de SS. Para os reprovados, no primeiro semestre de 2008, 17,3% dos alunos de Probabilidade e Estatística tiveram II e no segundo semestre 16,4% ficaram com MI.

A tabela 4.3 apresenta a distribuição dos alunos nas turmas. Pode-se perceber que Estatística Aplicada apresenta a maior quantidade de turmas, com Probabilidade e Estatística logo em seguida com 3 turmas a menos e Bioestatística com poucas turmas, 3 no primeiro semestre e 2 no segundo. Os alunos de bioestatística em 2008 só tiveram aulas com professores do quadro permanente, ao contrário de estatística aplicada cuja maioria, em torno de 65%, tiveram aula com professor substituto. Em PE a maior parte dos alunos também teve aula com professores do quadro permanente.

Tabela 4.3: Distribuição dos Estudantes nas Turmas

Variáveis		Disciplina						
		Bio		EA		PE		
		2008/1	2008/2	2008/1	2008/2	2008/1	2008/2	Geral
Turma	A	49,6%	60,4%	5,5%	6,6%	19,7%	19,1%	-
	B	-	-	12,9%	12,2%	19,1%	17,1%	-
	C	28,4%	-	9,7%	12,4%	12,4%	14,4%	-
	D	-	-	13,1%	12,6%	21,2%	25,2%	-
	E	22,0%	39,6%	11,2%	10,2%	16,7%	20,2%	-
	F	-	-	12,9%	12,9%	10,9%	-	-
	G	-	-	10,8%	12,4%	-	4,0%	-
	H	-	-	12,9%	11,4%	-	-	-
	I	-	-	11,0%	9,3%	-	-	-
Professor	Quadro	100%	100%	38,9%	30,3%	87,6%	100%	62,6%
	Substituto	-	-	61,1%	69,7%	12,4%	-	37,4%
Turno	Diurno	71,6%	100%	59,2%	61,4%	76,7%	81,5%	69,7%
	Noturno	-	-	27,7%	25,9%	-	-	14,4%
	Ambos	28,4%	-	13,1%	12,7%	23,3%	18,5%	15,9%
Horário	08:00 às 09:50	49,6%	60,4%	35,5%	37,6%	-	-	26,0%
	10:00 às 11:50	28,4%	-	23,9%	25,1%	19,1%	17,1%	21,2%
	14:00 às 15:50	22,0%	39,6%	12,9%	11,4%	37,9%	45,3%	24,4%
	16:00 às 17:50	-	-	-	-	19,7%	19,1%	6,7%
	19:00 às 20:50	-	-	16,7%	16,9%	-	-	9,0%
	20:50 às 22:40	-	-	11,0%	9,0%	23,3%	18,5%	12,7%
Local	Anfiteatro	-	-	65,3%	67,0%	52,7%	76,8%	57,8%
	Sala	100%	100%	34,7%	33,0%	47,2%	23,2%	42,2%

Em todas as disciplinas a maior parte dos alunos são de cursos cujo turno é diurno. Bioestatística e Probabilidade e Estatística tiveram alunos com curso em ambos turnos, enquanto que Estatística Aplicada apresenta uma melhor distribuição. Bioestatística e Probabilidade e Estatística não tiveram alunos com cursos noturnos, embora alguns alunos assistiram as aulas em horários noturnos, devido a serem de cursos com ambos turnos. Os alunos de Bioestatística só tiveram aulas em salas, enquanto as disciplinas Probabilidade e Estatística e Estatística Aplicada apresentaram uma maior diversidade, onde em todos os períodos mais da metade dos alunos tiveram aulas em anfiteatros.

4.3 Nível Turma

A tabela abaixo reflete em boa parte os resultados apresentados na tabela 4.3, que apresentava a distribuição dos alunos nas turmas.

Tabela 4.4: Perfil das Turmas e Professores

		Disciplina						
		Bio		EA		PE		
Variáveis		2008/1	2008/2	2008/1	2008/2	2008/1	2008/2	Geral
Professor	Quadro	100%	100%	44,4%	33,3%	83,3%	100%	65,7%
	Substituto	-	-	55,6%	66,7%	16,7%	-	34,3%
Turno	Diurno	66,7%	100%	55,6%	55,6%	66,7%	66,7%	62,9%
	Noturno	-	-	33,3%	33,3%	-	-	17,1%
	Ambos	33,3%	-	11,1%	11,1%	33,3%	33,3%	20,0%
Horário	08:00 às 09:50	33,4%	50,0%	33,4%	33,4%	-	-	22,9%
	10:00 às 11:50	33,3%	-	22,2%	22,2%	16,7%	16,7%	20,0%
	14:00 às 15:50	33,3%	50,0%	11,1%	11,1%	33,3%	33,3%	22,9%
	16:00 às 17:50	-	-	-	-	16,7%	16,7%	5,7%
	19:00 às 20:50	-	-	22,1%	22,2%	-	-	11,4%
	20:50 às 22:40	-	-	11,1%	11,1%	33,3%	33,3%	17,1%
Local	Anfiteatro	-	-	66,7%	66,7%	50,0%	66,7%	54,3%
	Sala	100%	100%	33,3%	33,3%	50,0%	33,3%	45,7%

Em bioestatística e no segundo período de Probabilidade e Estatística, todos os professores foram do quadro permanente, com melhor distribuição em estatística aplicada cuja maioria dos professores foram substitutos. A única disciplina que apresentou turmas para cursos noturnos é a de estatística aplicada. Nas demais disciplinas, a grande maioria das turmas eram para cursos diurnos. Além disso, pode-se perceber que mais de 60% de todos os alunos tiveram aulas entre 8 horas da manhã e 4 horas da tarde. Bioestatística só teve turmas em salas, enquanto os demais cursos tiveram mais turmas em anfiteatros, o que faz sentido devido a quantidade de alunos que cursam essas disciplinas.

4.4 Análise Bivariada

A análise bivariada permite verificar de forma exploratória quais variáveis influenciam na aprovação do aluno. Para as variáveis quantitativas ajustou-se um modelo de regressão logística simples para verificar a significância dessas variáveis. Para as variáveis qualitativas utilizou-se o teste qui-quadrado de associação, sendo consideradas na modelagem aquelas variáveis cujo p-valor está abaixo de 25%, pois decidiu-se ser mais liberal em uma análise inicial (veja Hosmer e Lemeshow 1989).

A tabela 4.5 apresenta os resultados para as variáveis quantitativas, para cada curso.

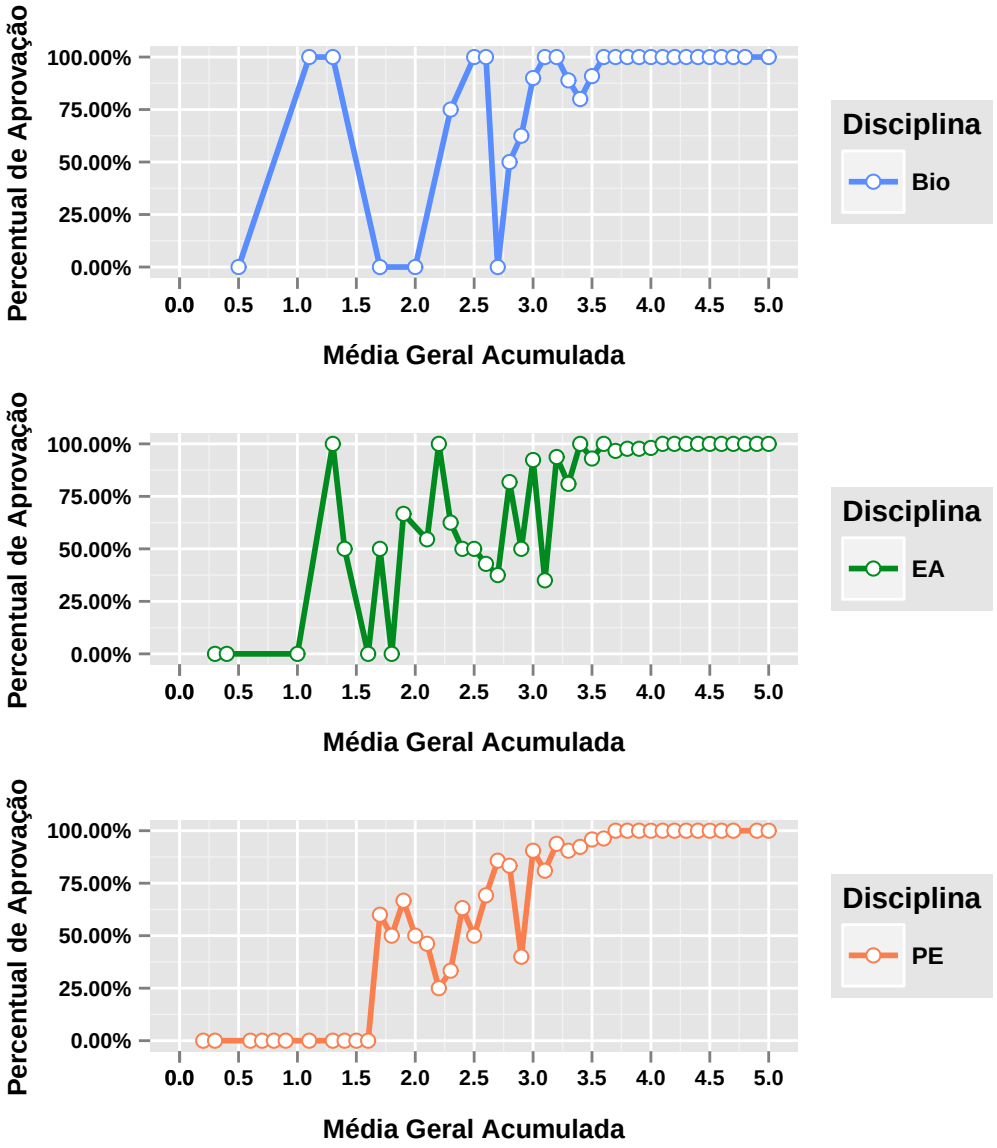
Tabela 4.5: Análise Bivariada das Variáveis Quantitativas

	Disciplina					
	Bioestatística		Estatística Aplicada		Probabilidade e Estatística	
Variáveis	Coeficiente	P-valor	Coeficiente	P-valor	Coeficiente	P-valor
MGA	1,99	$<$ 0,0001	2,21	$<$ 0,0001	2,38	$<$ 0,0001
Idade	-0,048	0,401	-0,014	0,651	0,028	0,425
Tempo(*)	-0,025	0,736	≈ 0	≈ 1	0,018	0,614

A única variável significativa é a MGA, em todas as disciplinas. com um coeficiente maior que 0, isso indica que quanto maior a MGA, maior a chance de aprovação do aluno. Assim, alunos que vão bem no semestre como um todo aparentam ter maior chance de ser aprovado nas disciplinas de serviço de estatística. Uma conclusão a respeito disso pode ser tirada na modelagem multinível, que inclui a variabilidade das turmas e a não independência dos dados.

O gráfico 4.6 mostra como foi o percentual de aprovação para cada valor da MGA.

Figura 4.6: Percentual de Aprovação por Disciplina para cada MGA



Para probabilidade e Estatísca temos um gráfico de dispersão com um comportamento muito próximo da curva em formato de S, isto é, o formato da regressão logística simples. Para estatística aplicada também aparenta se aproximar dessa curva, mas com variância maior. Bioestatística tem maior variância ainda. Vale ressaltar que nos boxplots 4.5 verificou-se que para bioestatística e EA, a maior parte dos dados está acima de 2, enquanto que para PE está

acima de 1. Dessa forma, alguns percentuais altos no gráfico de bioestatística para MGA baixo são referentes a alguns poucos alunos.

Assim, a variável MGA será considerada na modelagem em todas as disciplinas.

4.4.1 Bioestatística

A tabela 4.6 indica que a variável cotas foi significativa ao nível de 5%, sendo a variável sexo bem próxima de significativa. Em ambas variáveis temos uma certa diferença de percentual de aprovação, quase 10% a mais de aprovação para o sexo feminino e mais de 15% a mais de aprovação para alunos que não ingressaram por cotas. As variáveis que serão consideradas na modelagem são aquelas cujo p-valor na tabela 4.6 está abaixo de 25%. São elas: sexo, ano de ingresso, Modalidade, Semestre e Cotas.

Tabela 4.6: Percentual de Aprovação dos alunos de Bioestatística

Variáveis		Percentual de Aprovação	Estatística do Teste	P-valor
Sexo	Feminino	91,5%	3,7889	0,05159
	Masculino	82,6%		
País de Nascimento	Brasil	87,6%	≈ 0	≈ 1
	Exterior	100%		
Naturalidade	Distrito Federal	88,5%	4,0438	0,2568
	Goiás	77,8%		
	Minas Gerais	100%		
	Outras	83,3%		
UF de residência	Distrito Federal	88,0%	0,0413	0,839
	Outras	83,3%		
RA de residência	Brasília	87,5%	1,1223	0,5705
	Taguatinga	93,1%		
	Outras	85,2%		
Local de residência	DF Alta Renda	88,0%	1,4073	0,4948
	DF Média Renda	89,8%		
	Outros	80,8%		
Ano de Ingresso	2007	90,4%	3,621	0,1636
	2008	88,0%		
	Outros	80,0%		
Forma de Ingresso	PAS	91,8%	1,0481	0,306
	Outras	86,3%		
Curso	Agronomia	91,3%	5,0248	0,4129
	Ciências Biológicas	94,1%		
	Engenharia Florestal	82,6%		
	Farmácia	95,7%		
	Medicina Veterinária	87,0%		
	Outros	77,8%		
Modalidade	Obrigatória	86,0%	1,8641	0,1722
	Outras	93,5%		
Horário	08:00 - 09:50	89,6%	1,0302	0,5974
	10:00 - 11:50	87,1%		
	14:00 - 15:50	84,4%		
Semestre	2008/1	84,4%	2,1599	0,1417
	2008/2	91,1%		
Cotas	Sim	78,6%	6,2129	0,0127
	Não	94,6%		

Embora a variável curso tenha apresentado uma certa diferença de aprovação, como por exemplo, 82,6% dos alunos de Engenharia Florestal foram aprovados enquanto que esse percentual foi de 94,1% para os alunos de Ciências Biológicas, o p-valor de 0,4129 indica que essa diferença está longe de ser significativa.

Tabela 4.7: Aprovação em cada Turma de Bioestatística

Variáveis		Percentual de Aprovação	Estatística do Teste	P-valor
Turmas	A 2008/1	92,6%	18,8288	0,00085
	C 2008/1	87,1%		
	E 2008/1	62,5%		
	A 2008/2	86,9%		
	E 2008/2	97,5%		

A tabela 4.7 indica que há diferença significativa entre turmas, o que significa que faz sentido usar regressão multinível, visto que as turmas tem percentual de aprovação diferente. Algumas turmas apresentaram percentuais altos, em torno de 90%, enquanto a turma E em 2008/1 apresentou 62,5%, evidenciando que , de fato, as aprovações não são homogêneas entre as turmas.

4.4.2 Estatística Aplicada

A tabela 4.8 indica que as variáveis sexo, forma de ingresso, curso, professor, horário, turno e cotas foram significativas ao nível de 5%, com a variável local de residência bem próxima de ser significativa. Além dessas variáveis que foram altamente significativas, outras também serão consideradas na modelagem devido ao critério escolhido ser o p-valor abaixo de 25%, isto é, naturalidade, país de nascimento, ano de ingresso, modalidade e semestre.

Tabela 4.8: Percentual de Aprovação dos alunos de Estatística Aplicada

Variáveis		Percentual de Aprovação	Estatística do Teste	P-valor
Sexo	Feminino	90,8%	10,9395	$<$ 0,0001
	Masculino	83,7%		
País de Nascimento	Brasil	87,4%	1,4007	0,2366
	Exterior	76,2%		
Naturalidade	Distrito Federal	86,0%	7,8257	0,1661
	Goiás	86,4%		
	Minas Gerais	92,2%		
	Rio de Janeiro	87,5%		
	São Paulo	100%		
	Outras	90,6%		
UF de residência	Distrito Federal	87,2%	0,0417	0,8382
	Outras	84,4%		
RA de residência	Brasília	88,0%	2,9865	0,5601
	Taguatinga	86,0%		
	Sobradinho	90,9%		
	Guará	90,0%		
	Outros	84,1%		
Local de residência	DF Alta Renda	87,9%	7,2202	0,0652
	DF Média Renda	88,7%		
	DF Baixa Renda	78,8%		
	Outros	88,4%		
Ano de Ingresso	2005	89,3%	6,0063	0,1987
	2006	85,3%		
	2007	87,2%		
	2008	89,3%		
	Outros	76,3%		
Forma de Ingresso	Vestibular	86,7%	6,3271	0,04227
	PAS	91,1%		
	Outras	80,7%		
Curso	Administração	92,2%	30,0564	0,00043
	Arquivologia	72,4%		
	Biblioteconomia	89,0%		
	Ciência Política	90,1%		
	Ciências Contábeis	86,1%		
	Ciências Sociais	81,4%		
	Geografia	83,3%		
	Psicologia	94,5%		
	Relações Internacionais	91,6%		
	Outros	86,5%		
Modalidade	Obrigatória	87,4%	1,4648	0,2262
	Outras	81,8%		
Professor	Quadro	83,9%	4,8049	0,02838
	Substituto	88,8%		
Horário	08:00 - 09:50	88,4%	23,4355	0,00011
	10:00 - 11:50	92,4%		
	14:00 - 15:50	83,1%		
	19:00 - 20:40	77,3%		
	20:50 - 22:40	90,7%		
Local da Aula	Anfiteatro	87,6%	0,3162	0,5739
	Sala	86,3%		
Semestre	2008/1	85,8%	1,3717	0,2415
	2008/2	88,4%		
Turno	Ambos	98,4%	19,5839	$<$ 0,0001
	Diurno	86,9%		
	Noturno	82,3%		
Cotas	Sim	85,0%	5,4036	0,0201
	Não	92,0%		

Tabela 4.9: Aprovação em cada Turma de Estatística Aplicada

Variáveis		Percentual de Aprovação	Estatística do Teste	P-valor
Turmas	A 2008/1	88,4%	53,4102	$< 0{,}0001$
	B 2008/1	86,9%		
	C 2008/1	73,9%		
	D 2008/1	96,8%		
	E 2008/1	77,4%		
	F 2008/1	90,2%		
	G 2008/1	88,2%		
	H 2008/1	78,7%		
	I 2008/1	90,4%		
	A 2008/2	72,7%		
	B 2008/2	93,4%		
	C 2008/2	85,5%		
	D 2008/2	100%		
	E 2008/2	74,5%		
	F 2008/2	96,9%		
	G 2008/2	83,9%		
	H 2008/2	87,7%		
	I 2008/2	91,1%		

A tabela 4.9 indica que há diferença significativa entre turmas, o que significa que faz sentido usar regressão multinível, visto que as turmas tem percentual de aprovação diferente. De fato, algumas turmas apresentam percentuais em torno de 70% enquanto outras turmas estão em torno de 95%.

4.4.3 Probabilidade e Estatística

A tabela 4.10 aponta que as variáveis curso, horário e turno foram significativas ao nível de 5%. Além dessas variáveis que foram altamente significativas, outras também serão consideradas na modelagem devido ao critério escolhido ser o p-valor abaixo de 25%, isto é, sexo, professor e semestre.

Tabela 4.10: Percentual de Aprovação dos alunos de Probabilidade e Estatística

	Variáveis	Percentual de Aprovação	Estatística do Teste	P-valor
Sexo	Feminino Masculino	81,8% 75,7%	1,5589	0,2118
País de Nascimento	Brasil Exterior	76,3% 100%	1,0445	0,3068
Naturalidade	Distrito Federal Goiás Minas Gerais Rio de Janeiro São Paulo Outras	76,4% 77,9% 76,9% 68,0% 80,0% 75,6%	1,2724	0,9377
UF de residência	Distrito Federal Outras	76,5% 79,4%	0,1944	0,6593
RA de residência	Brasília Taguatinga Outros	76,7% 71,0% 78,1%	1,4889	0,475
Local de residência	DF Alta Renda DF Média Renda DF Baixa Renda Outros	77,3% 72,7% 83,3% 78,9%	2,6593	0,4472
Ano de Ingresso	2004 2005 2006 2007 2008 Outros	75,6% 84,1% 74,0% 73,0% 79,7% 85,7%	6,2423	0,2833
Forma de Ingresso	Vestibular PAS Outras	75,3% 82,2% 75,0%	2,553	0,279
Curso	Ciência da Computação Computação Engenharia Civil Engenharia de Redes de Comunicação Engenharia Elétrica Engenharia Mecânica Engenharia Mecatrônica Matemática Outros	86,1% 80,1% 66,1% 75,0% 83,8% 80,6% 79,2% 67,7% 85,2%	19,1667	0,01399
Modalidade	Obrigatória Outras	76,5% 77,1%	0,007	0,9334
Professor	Quadro Substituto	76,0% 85,4%	1,8832	0,17
Horário	10:00 - 11:50 14:00 - 15:50 19:00 - 20:40 20:50 - 22:40	69,3% 71,1% 85,2% 85,6%	18,75	0,00031
Local da Aula	Anfiteatro Sala	75,4% 78,7%	0,8415	0,359
Semestre	2008/1 2008/2	74,5% 78,9%	1,3935	0,2378
Turno	Ambos Diurno	85,6% 74,2%	6,9506	0,00838
Cotas	Sim Não	79,7% 82,9%	0,269	0,604

Tabela 4.11: Aprovação em cada Turma de Probabilidade e Estatística

Variáveis		Percentual de Aprovação	Estatística do Teste	P-valor
Turmas	A 2008/1	81,5%	33,6723	0,00041
	B 2008/1	63,5%		
	C 2008/1	85,4%		
	D 2008/1	81,4%		
	E 2008/1	58,2%		
	F 2008/1	80,5%		
	A 2008/2	89,5%		
	B 2008/2	76,5%		
	C 2008/2	86,0%		
	D 2008/2	70,7%		
	E 2008/2	71,7%		
	G 2008/2	100%		

A tabela 4.11 indica que há diferença significativa entre turmas, o que significa que faz sentido usar regressão multinível, visto que as turmas tem percentual de aprovação diferente. A diferença entre turmas parece maior ainda para essa disciplina onde algumas turmas tem percentual em torno de 60%, outras com 80% e uma única turma com 100% de aprovação, indicando que o professor possa ter maior impacto na aprovação dessa disciplina do que nas outras disciplinas.

Capítulo 5

Modelagem

Considerando as possíveis variáveis explicativas definidas na análise bivariada, iniciou-se a modelagem para cada uma das 3 disciplinas. Nessa etapa ajustou-se o modelo nulo para analisar o coeficiente de correlação intra-classe. A próxima etapa foi inserir uma variável explicativa por vez e verificando sua significância e o BIC do modelo. A variável explicativa que resultou no melhor modelo foi considerada na próxima etapa, na qual inseriu-se a segunda variável, repetindo os passos anteriores. Os modelos candidatos a modelo final foram considerados no diagnóstico. Finalmente escolheu-se o modelo e as inferências de interesse foram feitas.

Todos os resíduos e valores preditos utilizados foram do modelo condicional, isto é, que leva em conta a estimativa do efeito de cada turma por meio do BLUP (Best Linear Unbiased Predictor), visto que as turmas apresentam diferença de aprovação e é de interesse do estudo considerar essas diferenças. Além disso, testes de hipótese disponíveis no PROC GLIMMIX do SAS foram utilizados.

5.1 Estatística Aplicada

A categoria de sucesso nos modelos abaixo é aprovação. Primeiro ajustou-se o modelo nulo, para calcular o coeficiente de correlação intra-classe ρ e usar o teste de homogeneidade para testar as hipóteses:

H_0: Igualdade da matriz de variância e covariância de cada turma.
H_1: As matrizes de variância e covariância de cada turma são diferentes.

A tabela 5.1 apresenta os resultados desse ajuste.

Tabela 5.1: Modelo Nulo - Estatística Aplicada

Modelo	Valor	Erro Padrão	P-valor
Intercepto	1,9731	0,1859	< 0,0001
σ^2_G	0,4341	0,2291	0,0581
ρ	11,66%	-	-
Homogeneidade	11,42	-	0,8337

Tem-se que a variância entre turmas σ^2_G foi igual a 0,4341, resultando em um coeficiente de correlação intra-classe de 11,66%. O teste utilizado para verificar a significância dessa variância foi o teste de Wald com aproximação de Satterthwaite e com p-valor de 0,0581 será considerada como significativa. O teste de Homogeneidade indica que não se rejeita a hipótese de igualdade de variâncias em cada turma.

Após o ajuste do modelo nulo, inicia-se a etapa de seleção de modelos até que um ou vários candidatos a modelo final sejam obtidos. A tabela 5.2 apresenta os resultados do modelo final selecionado para Estatística Aplicada. As variáveis explicativas consideradas foram MGA e cotas e o modelo está representado abaixo:

$$logito(\pi_{ij}) = -5,8717 + 2,294MGA + 0,809COTAS + TURMA$$

Tabela 5.2: Modelo Final - Estatística Aplicada

Modelo	Valor	Erro Padrão	P-valor	Razão de Chances
Intercepto	-5,8717	0,8390	< 0,0001	-
MGA	2,2940	0,2347	< 0,0001	9,914
Cotas	0,8090	0,3655	0,0271	2,246
σ^2_G	0,8456	0,5093	-	-
Independência	9,96	-	0,0008	-

A variável MGA foi altamente significativa enquanto que cotas foi significativa com p-valor de 0,0271. A categoria de referência para a variável cotas é o aluno cotista, assim, a estimativa é para alunos que não passaram por cotas, em relação a alunos que passaram por cotas. O teste de Independência utlizado testa as seguintes hipóteses:

H_0: O modelo linear generalizado (MLG) se ajusta melhor do que o modelo linear generalizado misto (MLGM).

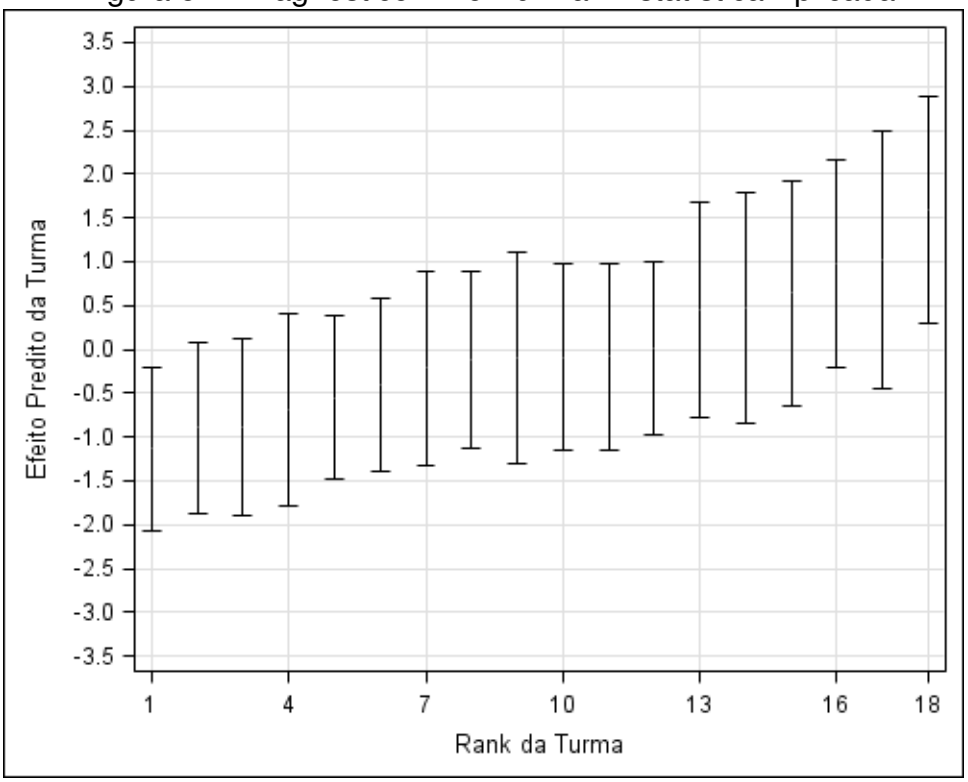
H_1: MLGM se ajusta melhor que MLG.

Ou seja, o teste é utilizado para verificar se há evidências de que o modelo que considera os alunos independentes (MLG) se ajusta melhor que o modelo que considera dependência entre alunos da mesma turma (MLGM). Outra forma de interpretação das hipóteses é que se a hipótese nula é rejeitada, existem evidências de que o modelo com efeito aleatório de turma se ajusta melhor do que o modelo sem esse efeito. Com p-valor de 0,0008 temos evidências de que o modelo MLGM se ajusta melhor, ou seja, a abordagem multinível produz um modelo com melhor ajuste.

Antes de interpretar os parâmetros do modelo e a razão de chances com seus respectivos intervalos de confiança, é necessário fazer o diagnóstico para verificar os pressupostos e valores discrepantes e influentes. O PROC GLIMMIX do SAS tem uma certa limitação nesse aspecto para regressão logística multinível. Pressuposto de normalidade para o efeito aleatório de turmas pode ser facilmente verificado, assim como a existência de valores discrepantes. Po-

rém não se verificou se esses valores discrepantes são influentes nas variâncias ou nas estimativas de β.

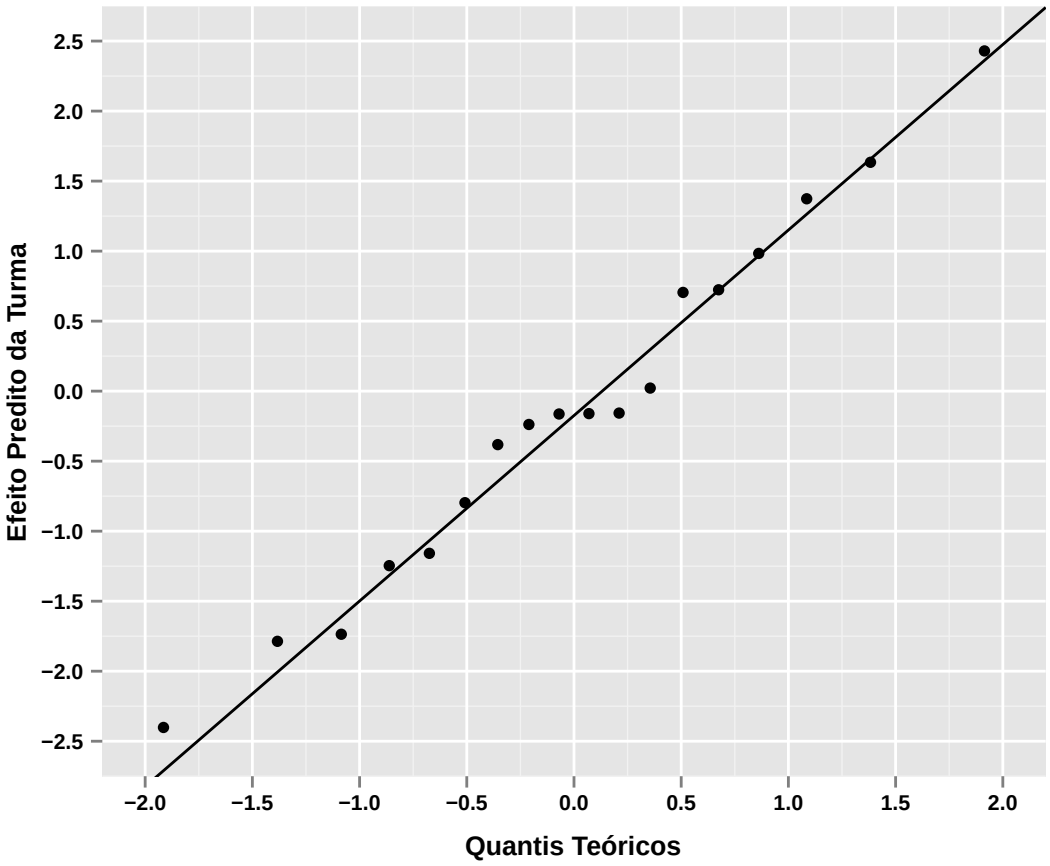
Figura 5.1: Diagnóstico Nível Turma - Estatística Aplicada



A figura 5.1 apresenta o efeito predito de cada turma pelo seu *rank*, isto é, ordenou-se do menor para o maior os efeitos preditos de cada turma, com barras do erro de predição. As turmas que interceptam zero não apresentam efeito significativamente diferente das demais, logo pode-se verificar que duas turmas, 1 e 18, apresentaram efeitos diferentes das outras.

Para verificar o pressuposto de normalidade do efeito aleatório da turma, basta verificar se os efeitos preditos estudentizados destas turmas são aproximadamente normais. A figura 5.2 apresenta o gráfico quantil-quantil , em relação aos quantis teóricos da distribuição normal, para essa medida.

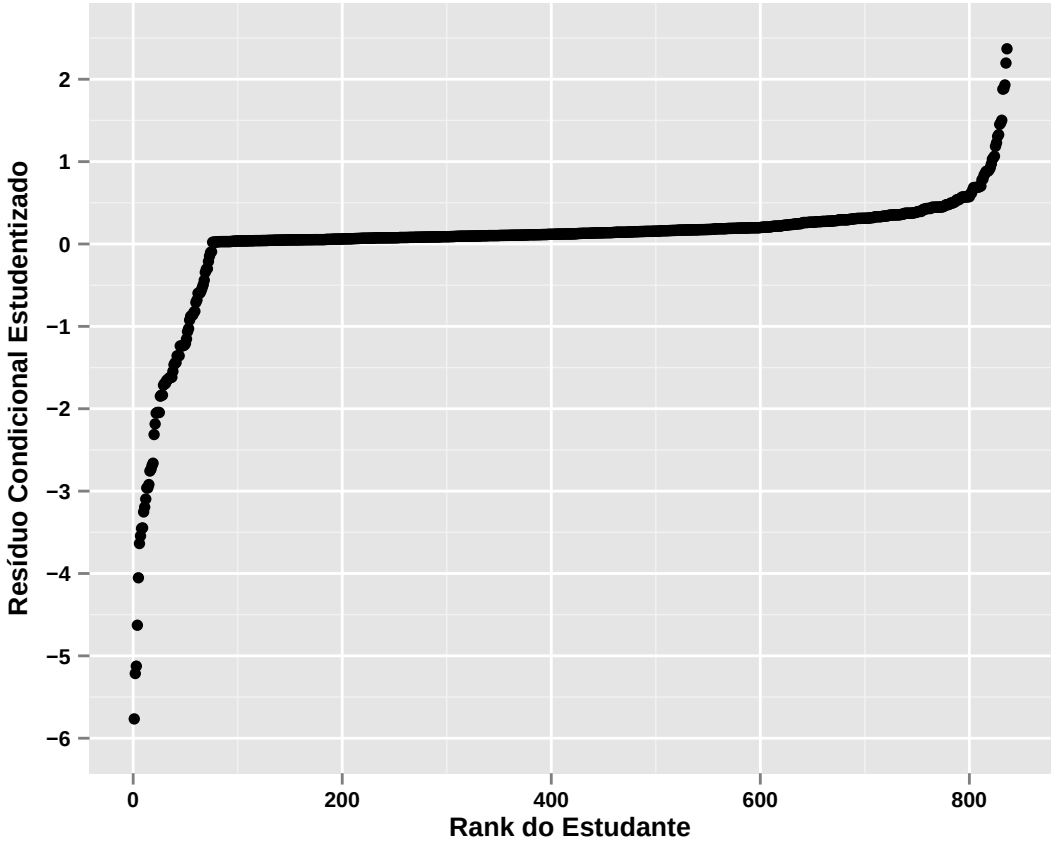
Figura 5.2: Gráfico Quantil-Quantil - Estatística Aplicada



Pela figura 5.2 temos evidências para não rejeitar o pressuposto de normalidade, mas isso pode ser questionado devido alguns desvios. O teste de normalidade de Shapiro-Wilk foi utilizado e o p-valor obtido foi 0,9629 o que indica que não se rejeita a hipótese nula de normalidade. Mais importante é observar que não se tem sérios desvios da normalidade no gráfico acima, então o modelo será considerado.

A figura 5.3 apresenta os resíduos estudentizados do nível aluno, ordenados de forma crescente.

Figura 5.3: Resíduos Estudentizados - Estatística Aplicada

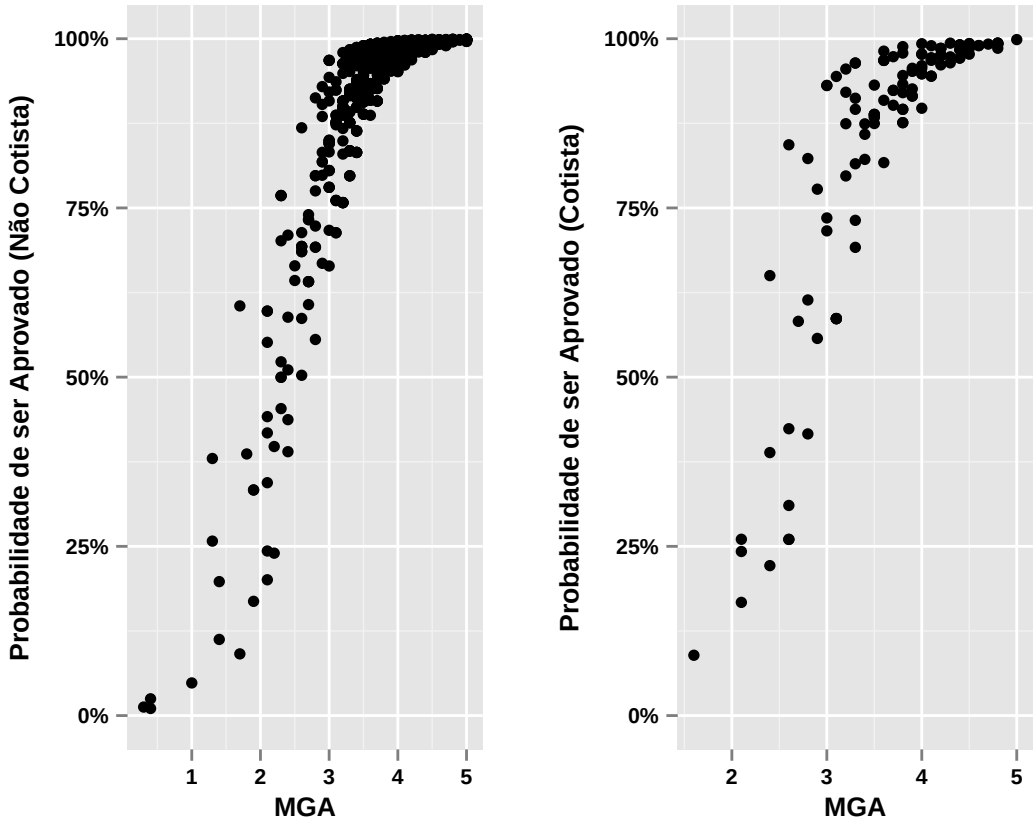


Embora a maior parte dos resíduos esteja concentrada em torno de -3 a 3, pode-se perceber a presença de alguns valores discrepantes, que podem ser devidos aos *outliers* presentes na variável MGA. Esses valores apresentaram resíduos negativos de até aproximadamente -6.

Em geral, o modelo parece ter se ajustado bem, mesmo com a presença de alguns valores discrepantes.

A figura 5.4 apresenta as probabilidades de aprovação preditas pelo modelo, variando os valores da MGA, para os cotistas e não cotistas.

Figura 5.4: Probabilidades Preditas - Estatística Aplicada



Como as probabilidades preditas levam em consideração os BLUP's, alunos com os mesmos valores das variáveis explicativas podem apresentar probabilidades preditas diferentes, dessa forma o gráfico apresentado é de dispersão, mas pode-se perceber claramente o formato de S da regressão logística. Se a regressão não fosse multinível, esse gráfico seria uma linha em formato de S. Pode-se perceber como os efeitos das turmas alteram a probabilidade de aprovação. Pode-se perceber também que o impacto na probabilidade de aprovação é maior quando se varia a MGA do que para cotistas e não cotistas, visto que ambas apresentam uma curvatura e probabilidades parecidas, embora para os cotistas ainda seja menor.

Voltando a tabela 5.2, como ambos coeficientes das variáveis explicativas são positivos, isso significa que quanto maior a MGA maior a chance de aprovação do aluno e que alunos não cotistas tem maior chance de aprovação do que alunos cotistas. Além disso, tem-se que a razão de chances para MGA é de 9,9, com intervalo de 95% confiança de 6,3 a 15,7. Para a variável cotas, tem-se que a razão de chances é de 2,2, com intervalo de 95% confiança variando de 1,1 a 4,6.

Assim, controlando cotas, para cada unidade da MGA, a chance de aprovação do estudante aumenta em 9,9 vezes. Por exemplo, a chance de um aluno com MGA 3 ser aprovado em Estatística Aplicada é aproximadamente 10 vezes a chance de um estudante com MGA 2 ser aprovado, controlando cotas. Analogamente, controlando MGA, a chance de um aluno não cotista ser aprovado é 2,2 vezes a chance de um aluno cotista ser aprovado.

5.2 Probabilidade e Estatística

A tabela 5.3 apresenta os resultados do nulo.

Tabela 5.3: Modelo Nulo - Probabilidade e Estatística

Modelo	Valor	Erro Padrão	P-valor
Intercepto	1,2753	0,1731	< 0,0001
σ^2_G	0,2320	0,1566	0,1366
ρ	6,6%	-	-
Homogeneidade	6,35	-	0,8490

Tem-se que a variância entre turmas σ^2_G foi igual a 0,232, resultando em um coeficiente de correlação intra-classe de 6,6%. Um valor relativamente baixo, mas continuou-se com a abordagem multinível. O p-valor do teste de Wald para σ^2_G foi 0,1366, que embora não seja significativa, continuou-se com a abordagem multinível, já que o p-valor não foi tão alto. O teste de Homogeneidade indica que não se rejeita a hipótese de igualdade de variâncias em cada turma.

A tabela 5.4 apresenta os resultados do modelo final selecionado para Probabilidade e Estatística. As variáveis explicativas consideradas foram MGA e Turno.

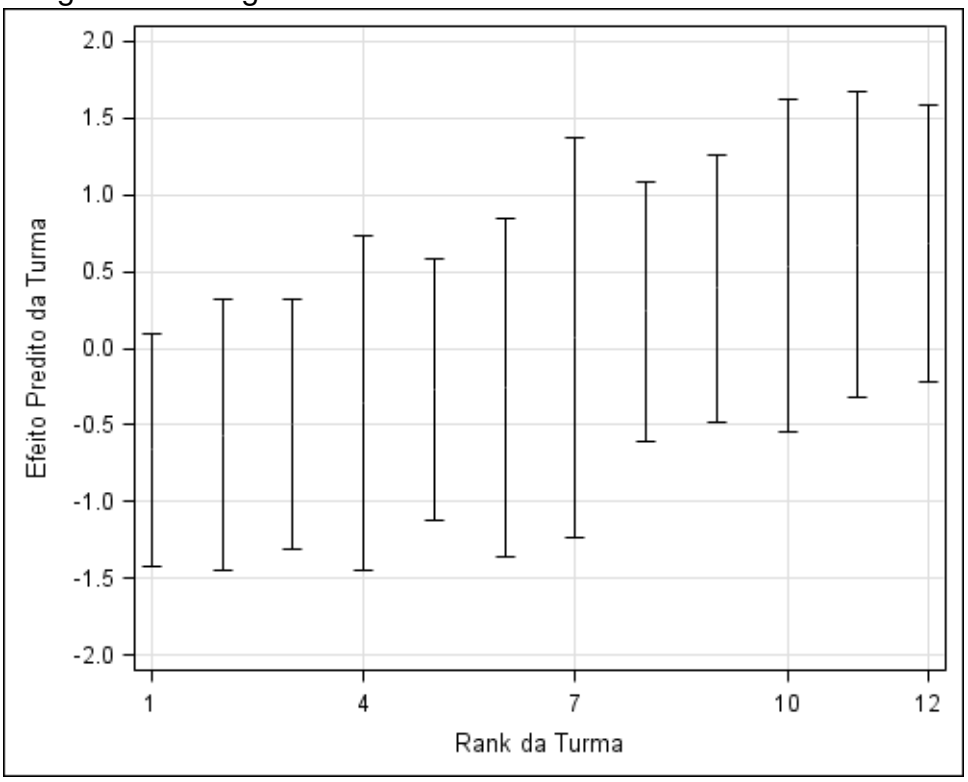
$$logito(\pi_{ij}) = -6,4097 + 2,6993MGA + 1,3762TURNO + TURMA$$

Tabela 5.4: Modelo Final - Probabilidade e Estatística

Modelo	Valor	Erro Padrão	P-valor	Razão de Chances
Intercepto	-6,4097	0,8225	< 0,0001	-
MGA	2,6993	0,2851	< 0,0001	14,870
Turno	1,3762	0,5940	0,0209	3,960
σ^2_G	0,3581	0,2865	-	-
Independência	5,55	-	0,0093	-

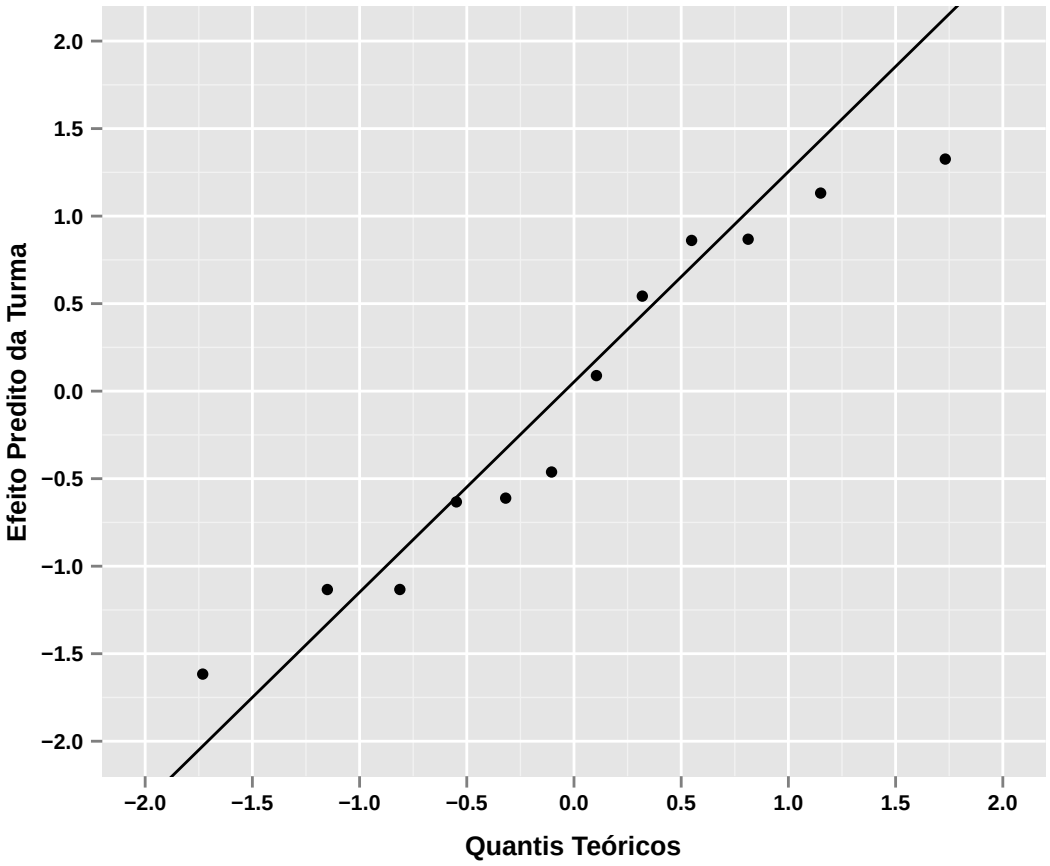
A variável MGA foi altamente significativa enquanto que turno foi significativo com p-valor de 0,0209. A categoria de referência para a variável turno é Diurno, assim, a estimativa é para alunos cujo curso é de ambos turnos, em relação a alunos cujo curso é diurno. O teste de Independência foi utilizado e o p-valor de 0,0093 indica que temos evidências de que o modelo MLGM se ajusta melhor, ou seja, a abordagem multinível produz um modelo com melhor ajuste, embora não se tenha tanta variabilidade entre turmas.

Figura 5.5: Diagnóstico Nível Turma - Probabilidade e Estatística



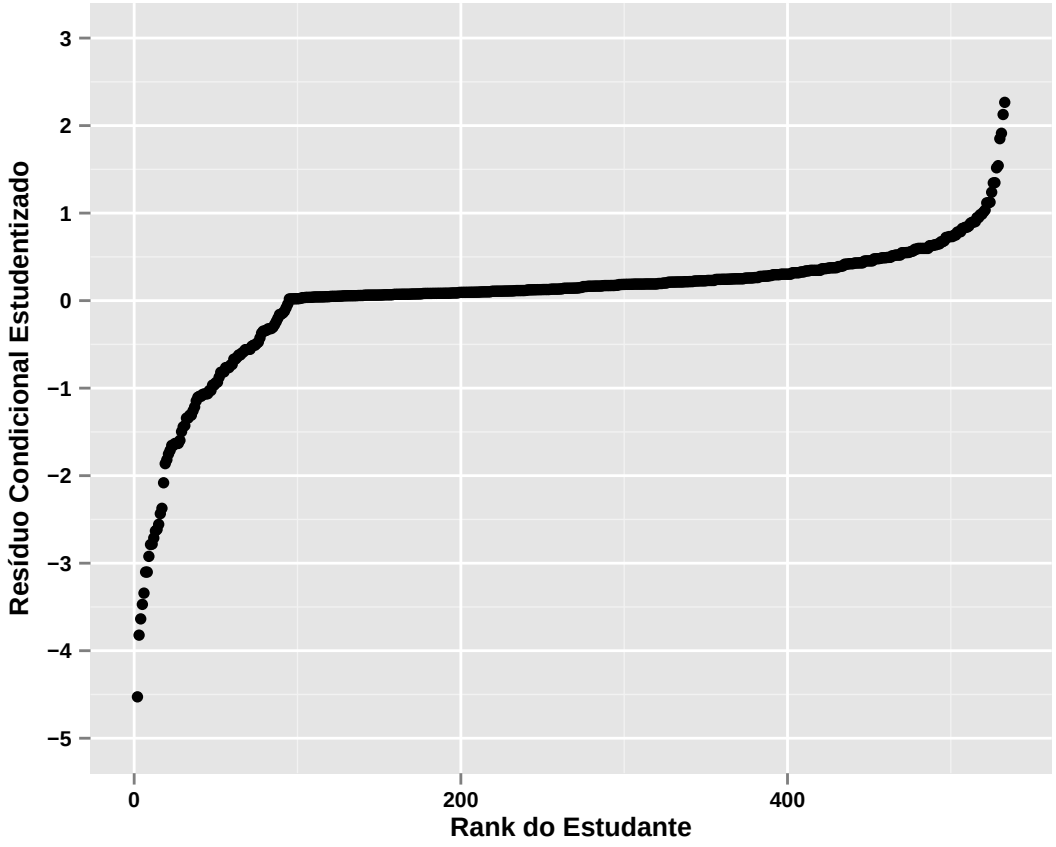
A figura 5.5 apresenta o efeito predito de cada turma, pelo rank da turma, com barras do erro de predição. As turmas que interceptam zero não apresentam efeito significativamente diferente das demais, logo pode-se verificar que embora se tenha alguma diferença, todas turmas interceptam zero. Acreditando que o modelo multinível ainda assim apresenta um melhor ajuste, continuou-se com esta abordagem.

Figura 5.6: Gráfico Quantil-Quantil - Probabilidade e Estatística



No gráfico quantil-quantil apresentado na figura 5.6 tem-se que o pressuposto de normalidade não parece ser aceito, com alguns problemas nos extremos, indicando caudas pesadas ou até turmas com efeitos discrepantes. O teste de normalidade de Shapiro-Wilk foi utilizado e o p-valor obtido foi 0,3933 o que indica que não se rejeita a hipótese nula de normalidade, mas pelo gráfico quantil-quantil, a normalidade aqui é duvidosa.

Figura 5.7: Resíduos Estudentizados - Probabilidade e Estatística

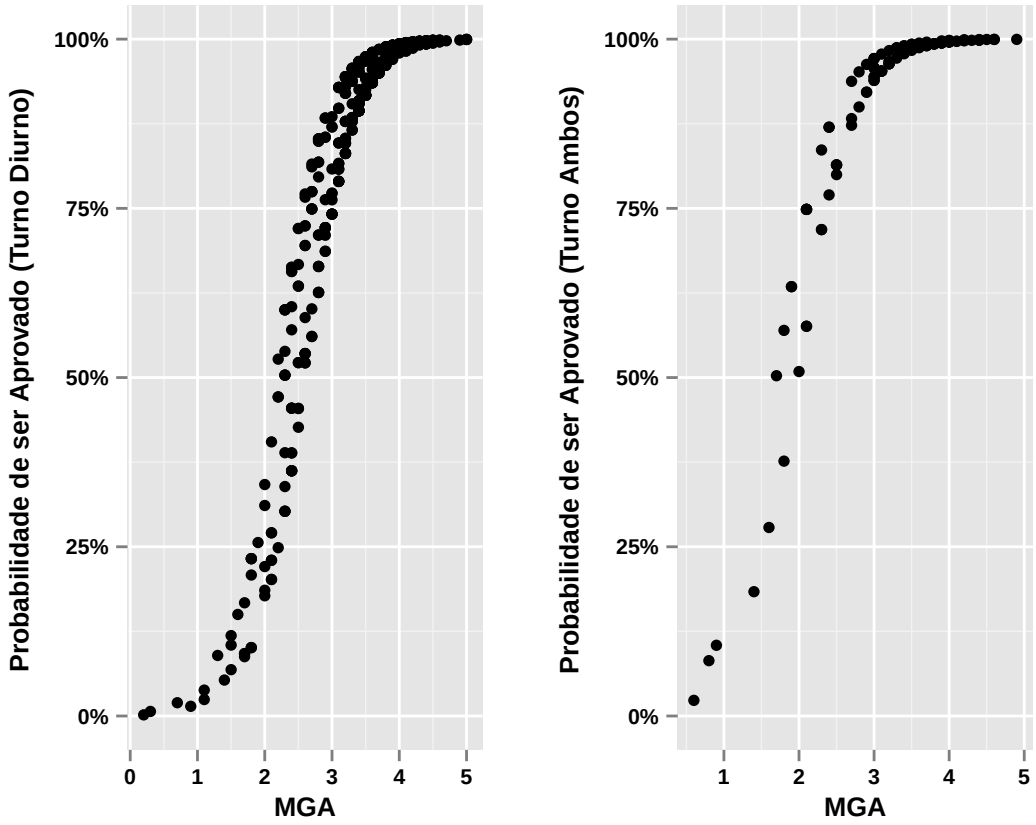


No gráfico 5.7 observa-se que embora a maior parte dos resíduos esteja concentrada em torno de -3 a 3, pode-se perceber a presença de alguns valores discrepantes, que podem ser devidos aos *outliers* presentes na variável MGA. Esses valores apresentaram resíduos negativos de até aproximadamente -5.

Em geral, o modelo não parece ter se ajustado bem principalmente pelo fato dos efeitos das turmas não serem próximos da distribuição normal. O baixo coeficiente de correlação intra-classe e pouca diferença dos efeitos entre as turmas indicam que talvez o modelo sem efeito aleatório de turma possa ser melhor. Entretanto um dos testes utilizados indicou que o ajuste do modelo é melhor com efeitos aleatórios de turma. Mas o mais importante é que os

pressupostos do modelo sejam aproximadamente satisfeitos, o que não foi o caso.

Figura 5.8: Probabilidades Preditas - Probabilidade e Estatística



Na figura 5.8 observa-se que, ao contrário da disciplina Estatística Aplicada, a disciplina Probabilidade e Estatística apresentou probabilidades preditas bem próximas em diferentes turmas, o que vai de encontro com o resultado do baixo coeficiente de correlação intra-classe. Pode-se perceber também que o impacto na probabilidade de aprovação é maior quando se varia a MGA acumulada do que para a variável turno, visto que ambas apresentam uma curvatura e probabilidades parecidas, embora para alunos cujo curso é de ambos turnos seja maior.

Assim, a abordagem multinível não será considerada para essa disciplina, pois o pressuposto de normalidade é duvidoso e não parece haver muita diferença entre turmas. Neste caso deve-se utilizar o modelo de regressão logística múltipla, desconsiderando efeito aleatório de turmas.

5.3 Bioestatística

A tabela 5.5 apresenta os resultados do modelo nulo para bioestatística.

Tabela 5.5: Modelo Nulo - Bioestatística

Modelo	Valor	Erro Padrão	P-valor
Intercepto	1,9785	0,4614	0,0128
σ^2_G	0,8034	0,7755	0,3002
ρ	19,63	-	-
Homogeneidade	3,19	-	0,5272

Tem-se que a variância entre turmas σ^2_G foi igual a 0,8034, resultando em um coeficiente de correlação intra-classe de 19,63$\%$. O teste de Wald para esse componente da variância resultou em um p-valor de 0,3002, que é um reflexo do erro padrão de 0,7755, indicando que não se tem variabilidade significativa entre turmas. O teste de Homogeneidade indica que não se rejeita a hipótese de igualdade de variâncias em cada turma.

A tabela 5.6 apresenta os resultados do modelo final selecionado para Bioestatística. As variáveis explicativas consideradas foram MGA e cotas.

Tabela 5.6: Modelo Final - Bioestatística

Modelo	Valor	Erro Padrão	P-valor	Razão de Chances
Intercepto	-7,7082	2,6103	0,0418	-
MGA	3,1568	0,9257	0,0008	23,495
Cotas	1,6526	0,8072	0,0422	5,221
σ^2_G	3,9660	4,6935	-	-
Independência	7,42	-	0,0032	-

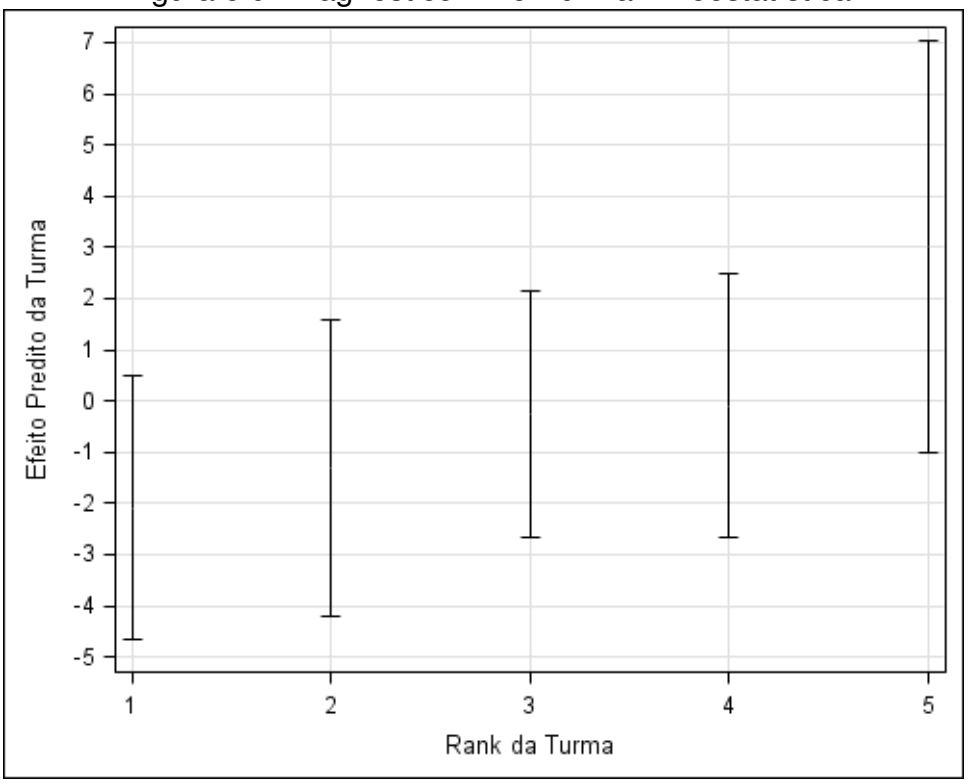
A variável MGA foi altamente significativa enquanto que cotas foi significativa com p-valor de 0,0422. A categoria de referência para a variável cotas é o

aluno cotista, assim, a estimativa é para alunos não cotistas, em relação a alunos cotistas. O teste de Independência foi utilizado e o p-valor de 0,0032 indica que temos evidências de que o modelo MLGM se ajusta melhor. Entretanto, observa-se um alto erro padrão em todas as estimativas, então esse modelo não parece ter um bom ajuste e pode-se suspeitar de *overdispersion*, variância muito alta do que se esperaria se o modelo fosse próximo da realidade.

Tem-se, por exemplo, que a razão de chances para MGA é de 23,495, com um intervalo de 95% muito grande: de 3,778 a 146,090. Algo está claramente errado com o modelo. A razão de chances para cotas também é muito alta: de 1,061 a 25,692.

Antes de se descartar o modelo, verificou-se por meio do diagnóstico o que poderia estar causando esse problema.

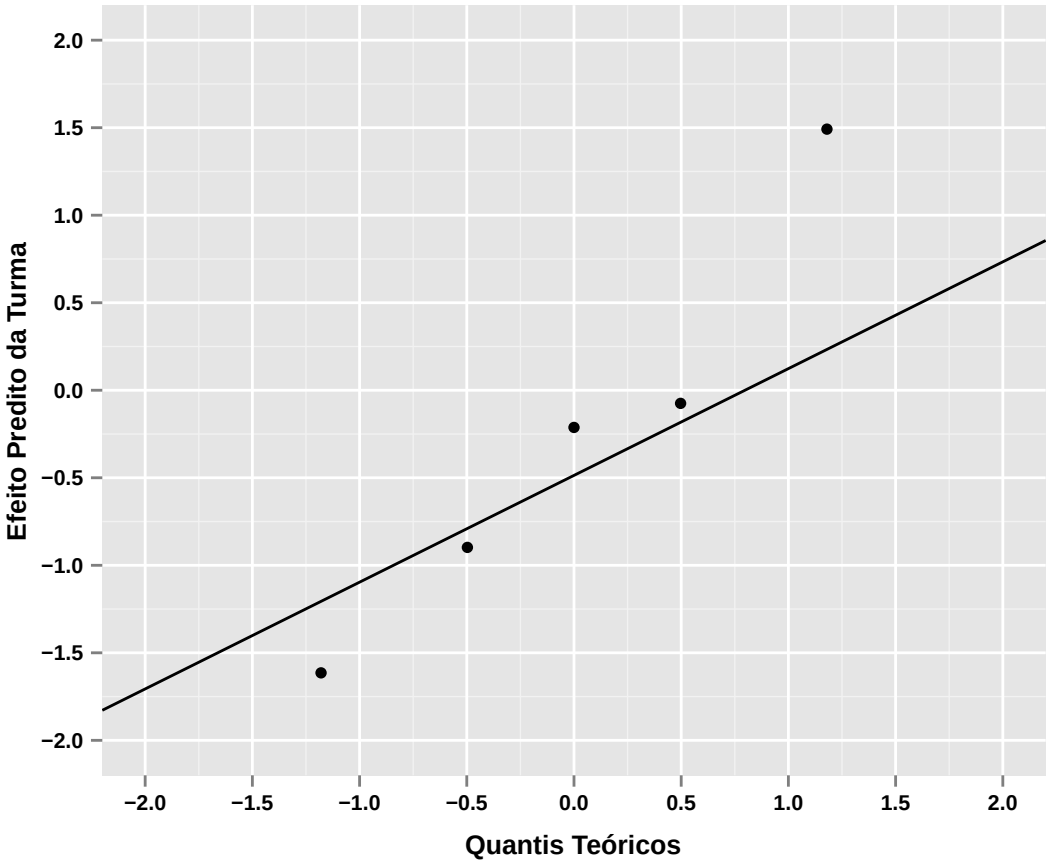
Figura 5.9: Diagnóstico Nível Turma - Bioestatística



Na figura 5.9 tem-se o efeito predito de cada turma, com barras do erro de

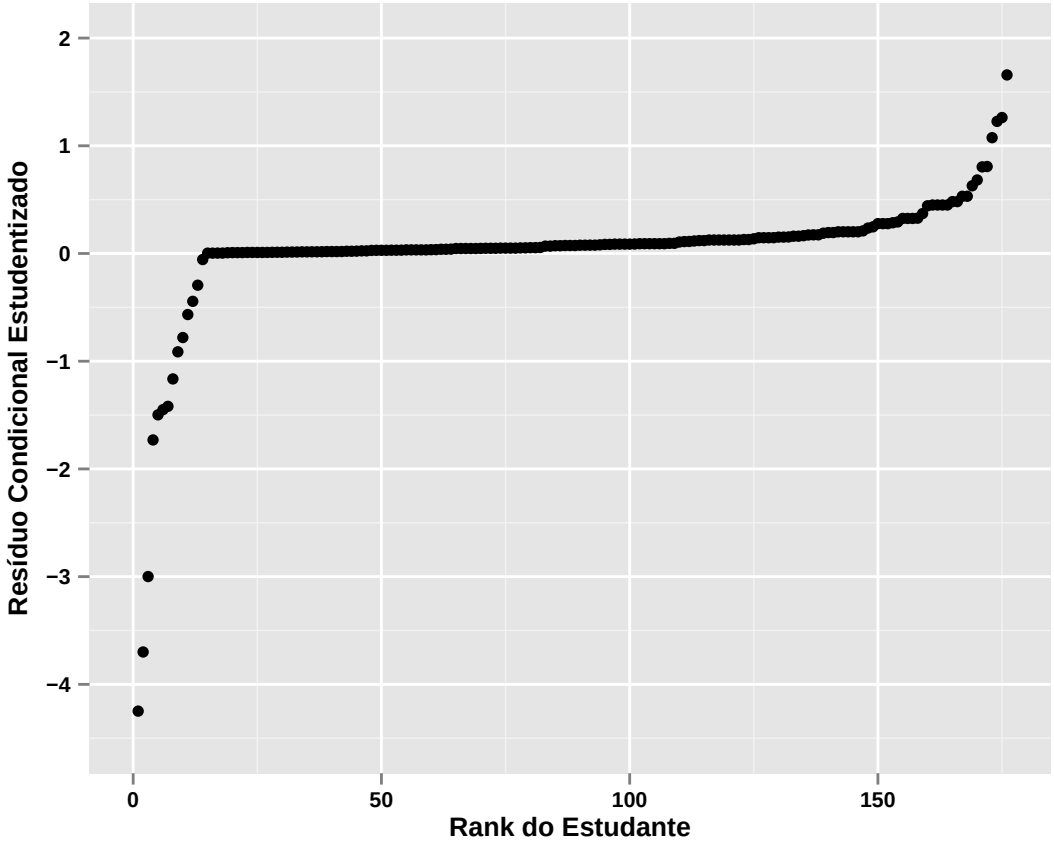
predição. Tem-se que todas as turmas interceptam 0, assim o efeito de cada uma delas não difere significativamente da outra, que era esperado devido a não significância da variabilidade entre turmas.

Figura 5.10: Gráfico Quantil-Quantil - Bioestatística



Na figura 5.10 não se tem evidências de normalidade mesmo com uma quantidade pequena de turmas, pois além de nenhum dos pontos está próximo da reta e um deles está extremamente distante.

Figura 5.11: Resíduos Estudentizados - Bioestatística

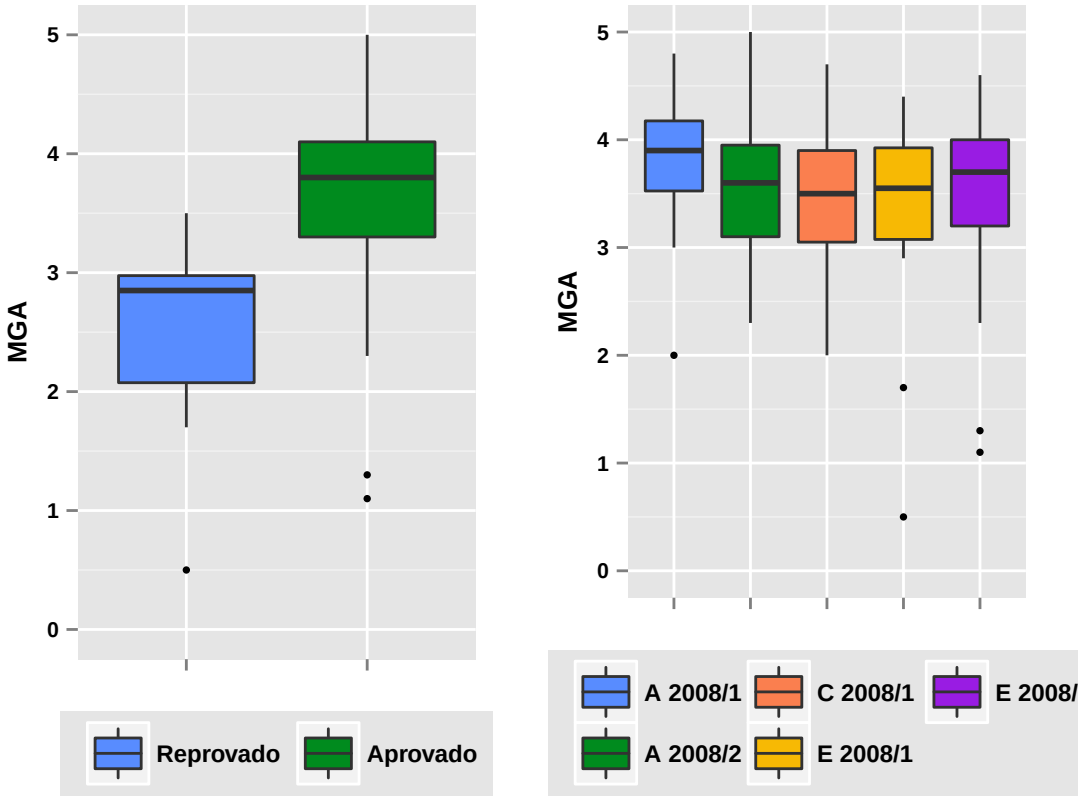


A figura abaixo 5.11 apresenta os resíduos estudentizados do nível aluno.Embora a maior parte dos resíduos esteja concentrada em torno de -3 a 3, pode-se perceber a presença de alguns valores discrepantes, que podem ser devidos aos valores discrepantes presentes na variável MGA. Esses valores apresentaram resíduos negativos de até aproximadamente -4. Esta disciplina apresenta menos valores discrepantes do que as disciplinas Estatística Aplicada e Probabilidade e Estatística. Isso indica que provavelmente, o problema com esse modelo não é de valores discrepantes.

Para reforçar a idéia de que o problema do modelo não é de valores discrepantes, fez-se um gráfico dos *boxplots* da variável explicativa MGA por aprova-

ção e por turmas, apresentado na figura 5.12

Figura 5.12: *Boxplot* da MGA por Aprovação e Turmas - Bioestatística



Assim, não se observa grande variabilidade nem muitos valores discrepantes. A grande variabilidade do modelo pode ser então devido a uma das turmas ser discrepante em relação as outras. Na análise descritiva verificou-se que a turma E de 2008/1 tinha 62,5%, muito diferente das outras turmas que tinham em torno de 90%. Justamente essa turma, representada pelo rank 5, apresentou comportamento diferente das demais nas figuras 5.9 e 5.10, com maior variabilidade.

Assim, os valores grandes de erro padrão devem ser devido a má especificação do modelo, mesmo que o teste da independência tenha rejeitado a

hipótese de que o modelo sem o efeito aleatório de turmas é significativamente melhor. Retirando o efeito aleatório e ajustando um modelo de regressão logística múltipla, obteve-se estimativas com erro padrão aceitável. Esse modelo está apresentado na tabela 5.7.

Tabela 5.7: Regressão Logística Múltipla - Bioestatística

Modelo	Valor	Erro Padrão	P-valor	Razão de Chances
Intercepto	-3,5400	1,4056	0,0118	-
MGA	2,0164	0,4651	< 0,0001	7,5113
Cotas	1,6675	0,6862	0,0151	5,2989

Aqui, os intervalos de 95% confiança para MGA e cotas são, respectivamente, de 3,01 a 18,69 e de 1,38 a 20,34. Um resultado não muito diferente das estimativas de cotas para o modelo multinível, mas um intervalo muito menor para a variável MGA. Assim, não considerar os efeitos aleatórios de turma, para essa disciplina, foi a melhor opção. Como o interesse do trabalho é a regressão logística multinível, esse modelo não será analisado, apenas mencionado como uma solução para o problema no ajuste do modelo multinível.

Capítulo 6

Conclusão

A abordagem de modelos mistos é muito abrangente, incluindo vários tipos de problemas, como no caso da estrutura dos dados na aplicação apresentada. Com todas as análises feitas, conclui-se que essa abordagem para regressão hierárquica pode apresentar ganhos significativos nas estimativas dos efeitos desejados e das inferências feitas.

Para a disciplina Estatística Aplicada, 11,68$\%$ da variância no desempenho dos alunos pode ser atribuída a turma em que se matricularam. Para Bioestatística, a abordagem multinível não apresentou um bom ajuste, de forma que para se ajustar um modelo com as variáveis explicativas MGA e cotas, será necessária a abordagem usual, então essa disciplina foi desconsiderada de uma análise mais profunda. Para Probabilidade e Estatística, a abordagem multinível não foi necessária e nem se obteve um bom ajuste. Vale ressaltar que os dados são observacionais, então problemas com os pressupostos do modelo e observações discrepantes são esperados.

A variável Média Geral Acumulada (MGA) foi altamente significativa na explicação da aprovação do aluno em Estatística Aplicada, de forma que quanto maior o MGA do aluno maior a chance dele ser aprovado, indicando que o desempenho dos alunos dessa disciplina depende de seu desempenho geral

naquele semestre.

Outras variáveis foram consideradas significativas isoladamente, como forma de ingresso, sexo, curso e professor, mas com a inclusão e a alta significância da variável MGA, essas variáveis foram desconsideradas. Isso significa que outros fatores afetam o rendimento do aluno, mas que MGA é a principal variável responsável pela aprovação desse aluno.

Para Estatística Aplicada, a variável cotas teve efeito significativo, indicando que alunos que ingressaram na UnB por meio do sistema de cotas tiveram menor chance de serem aprovados nessa disciplina.

Como a abordagem de modelos mistos é mais complexa, problemas nos algoritmos de estimação e limitação na análise de diagnóstico são frequentemente encontrados. Não houve problemas na estimação com o uso do SAS, mas as possibilidades de diagnóstico foram limitadas. Contudo, o SAS permite que o usuário programe aquilo que não se tem em seus procedimentos, mas é necessário um conhecimento mais profundo do modelo e demanda um tempo maior de estudo.

Esse trabalho se limitou ao ano de 2008, então recomenda-se que novos trabalhos sejam feitos considerando anos mais atuais, ou um período de anos para se fazer comparações. Uma outra sugestão é um estudo de diagnóstico de influência em regressão logística multinível.

Referências Bibliográficas

[1] AGRESTI, Alan. **An Introduction to Categorical Data Analysis**. Second Edition. Hoboken, New Jersey: John Wiley & Sons, 2007.

[2] AKAIKE, Hirotugu. A New Look at the Statistical Model Identification. **IEEE Transaction on Automatic Control**. 1974; 19 (6), 716–723.

[3] DAI, Jian; LI, Zhongmin; DAVID, Rocke. **Hierarchical Logistic Regression Modeling with SAS GLIMMIX**. University of California, 2006.

[4] de LEEUW, Jan. **Random Coefficient Models for Multilevel Analysis**. Department of Statistics, UCLA, 2006.

[5] DEMIDENKO, Eugene. **Mixed Models: Theory and Applications**. Hoboken, New Jersey: John Wiley & Sons, 2004.

[6] FERRAZ, Amanda. Avaliação do rendimento dos alunos em disciplinas ofertadas pelo departamento de estatística para outros cursos da universidade de brasília: uma aplicação de regressão logística multinível [**Trabalho de conclusão de curso**] . Brasília, Distrito Federal: Universidade de Brasília, 2013.

[7] HASTINGS, Nicholas; PEACOCK, Brian; EVANS, Merran; FORBES, Catherine. **Statistical Distributions**. Fourth Edition. Hoboken, New Jersey: John Wiley & Sons, 2011.

[8] HOSMER, David; LEMESHOW, Stanley. **Applied Logistic Regression**, United States: John Wiley & Sons,1989.

[9] HOX, J. J.. **Multilevel analysis: techniques and applications**. Second Edition. Great Britain: Routledge, 2010.

[10] JUNCO, Reynol. Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance.**Computers in Human Behavior**. Department of Academic Development and Counseling, Lock Haven University, 104 Russell Hall, Lock Haven, PA 17745, United States. 2011.

[11] LAROS, Jacob; MARCIANO, João. Análise multinível aplicada aos dados do nels:88. **Estudos em avaliação educacional**. 2008 Maio; 19 (40) : 263-278.

[12] LAROS, Jacob; MARCIANO, João; ANDRADE, Josemberg. Fatores que Afetam o Desempenho na Prova de Matemática do SAEB: um Estudo Multinível. **Avaliação Psicológica**. 2010; 9 (2) : 173-186.

[13] LI, Jia; ALTERMAN, Toni; DEDDENS, James A. Analysis of Large Hierarchical Data with Multilevel Logistic Modeling using SAS PROC GLIMMIX. **Proceedings of the Thirty-first Annual SAS® Users Group International Conference**. Cary, NC: SAS Institute Inc., 2006.

[14] LITTEL, Ramon C.; MILLIKEN, George; STROUP, Walter; WOLFINGER, Russel; SCHABENBERGER, Oliver. **SAS FOR MIXED MODELS**. Second Edition. Cary, NC: SAS Institute Inc., 2006.

[15] PREGIBON, Daryl. Logistic Regression Diagnostics. **The Annals of Statistics**. 1981; 9 (4) : 705-724.

[16] SEARLE, S. R.. **Linear Models**. New York: John Wiley & Sons, 1971.

[17] SEARLE, S.R.; CASELLA, G.; McCULLOCH, C.E. **Variance Components**. Hoboken, New Jersey: John Wiley & Sons, 2006.

[18] SCHWARZ, Gideon. Estimating the Dimension of a Model. **Annals of Statistics**. 1978; 6 (2), 461–464.

[19] SINGER, J. D.. Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. **Journal of Educational and Behavioral Statistics**. 1998; 24 (4) : 323-355.

Apêndice

O PROC GLIMMIX do SAS é o procedimento utilizado para análise de modelos lineares generalizados mistos, incluindo regressão logística multinível. Esse procedimento apresenta algumas limitações na análise de resíduos para o caso variável binária e ligação logito, no sentido de ter menos recursos que o PROC LOGISTIC ou PROC GENMOD.

Toda a programação apresentada aqui se refere ao modelo de Estatística Aplicada. A variável resposta aprovado ou reprovado foi codificada da forma 1 e 0, respectivamente. O modelo nulo pode ser ajustado da seguinte forma:

PROC GLIMMIX data=dados;
class turma;
model aprovacao(event='1') = / dist=binary link=logit ddfm=bw solution;
random turma / solution cl;
run;

O comando *class* é utilizado para identificar quais as variáveis categóricas que estão sendo utilizadas. A opção *event='1'* permite especificar qual a probabilidade de sucesso que está sendo modelada. Os comandos *dist=binary* e *link=logit* especificam qual o modelo que está sendo ajustado, isto é, o modelo de regressão logística multinível. A opção *ddfm=bw* especifica qual o método utilizado para calcular os graus de liberdade para os testes dos efeitos fixos, enquanto que *solution* pede a solução dos efeitos fixos. O comando *random* especifica os efeitos aleatórios do modelo, sendo nesse caso apenas o efeito

da turma, enquanto que as opções desse comando *solution* e *cl* pedem, respectivamente, as soluções dos efeitos de cada turma (BLUP) e os intervalos de confiança.

Para se fazer o teste de homogeneidade das variâncias em cada turma, deve-se usar:

random turma / group=turma;
covtest homogeneity;

No comando *random* a opção *group* indica que cada turma tem sua própria variância que deve ser estimada. O comando *covtest* especifica qual o teste que será feito, no caso o teste de homogeneidade, onde tem-se como hipótese nula H_0 : As variâncias de cada turma são iguais.

Para testar a significância da variância e obter os intervalos de confiança deve-se usar:

covtest / wald cl(type=elr);

O comando *covtest* aparece com as opções *wald* e *cl(type=elr)*, que são respectivamente, o teste de Wald para a significância da variância entre turmas e o tipo do intervalo de confiança pedido.

Já o modelo com as variáveis explicativas pode ser ajustado da seguinte forma:

PROC GLIMMIX data=dados method=laplace;
class turma cotas;
model aprovacao(event='1') = MGA cotas / dist=binary link=logit ddfm=bw
solution oddsratio;
random turma / solution cl;
output out=saida pred=p stderr=stdlp resid=r student=s lcl=lplcl ucl=lpucl
pearson=rp;
covtest glm;
run;

Nesse modelo poucos comandos são diferentes do modelo nulo. A opção *method=laplace* do PROC GLIMMIX permite especificar qual o método de utilizado para se estimar os parâmetros, onde o método default do SAS é *Residual Pseudo-Likelihood*. A opção *oddsratio* no comando model fornece as estimativas da razão de chances com intervalos de confiança. O comando *output out=saida* pede para guardar as saídas desejadas em um banco de dados chamado saida. Pode-se pedir os valores preditos do modelo (com ou sem BLUP), intervalos de confiança, vários tipos de resíduos entre outros. Finalmente, a opção *covtest glm* pede para testar a hipótese de que o modelo linear generalizado misto (com dependência de alunos na mesma turma) se ajusta significativamente melhor que o linear generalizado (todos os alunos independentes), onde a hipótese nula é H_0 : MLG se ajusta tão bem quanto o MLGM.

Para se obter as estimativas (BLUP) dos efeitos de cada turma em uma saída, basta pedir no output *solutionr*, como mostrado no comando abaixo:

ods output solutionr=solucoes;